Published in AJSE, Vol: 25, Issue:2
Received on 5th August 2026
Revised on 26th September 2026
Published on 31st October 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/sa1d7q88

Can Synthetic Data Replace Real Data? Evaluating the Efficacy, Limits, and Label Trustworthiness of Generative Data in Supervised Learning

Niyaz Ahmad Khan, Md Yousuf Bin Nazib, Eshrath Jahan Esha, Golam Mostofa Naeem, Islam Uddin, and Abu Ubayda

***Abstract*—Synthetic data are increasingly proposed as substitutes for real training records where privacy restricts data sharing. We evaluate this substitution on a medical appointment no-show dataset of 110,526 records with the Train-Synthetic-Test-Real protocol, five classifiers, six synthetic mixing ratios and nine generation runs of CTGAN, TVAE and a Gaussian copula on this and a second dataset (UCI Adult). Full substitution lowered mean AUC-ROC in every run (by 5.72% to 11.46% across classifiers in the primary run), and mixed training beat training on its real part alone in at most six of 40 paired DeLong comparisons per run (Holm correction) while losing in 6 to 30. We propose Synthetic Label Trustworthiness Estimation (SLTE), an empirical label-consistency check that scores each synthetic label against a model trained on real data, relabels low-confidence samples by nearest-neighbour vote and down-weights uncertain ones. SLTE changed 12.88% of the primary synthetic labels, 94.2% of them from minority to majority class, but a label-permutation null reproduced this asymmetry and the same procedure changed 14.20% of held-out real labels, so the changes reflect disagreement with a real-data model rather than generator errors. In the primary run SLTE improved AUC-ROC significantly only at 80% and 100% synthetic data, recovering 57.8% of Gradient Boosting's loss at full substitution, but it lowered recall for most classifiers, and in other runs it lowered AUC-ROC for Decision Tree, Logistic Regression or Naive Bayes. Synthetic data did not replace real data here; SLTE is a quality-control step for when substitution cannot be avoided.**

***Index Terms*—Class imbalance, CTGAN, healthcare machine learning, label corruption, SLTE framework, supervised learning, synthetic data, TSTR protocol.**

I. INTRODUCTION

IN health care, regulations such as HIPAA and GDPR restrict sharing the labelled records that supervised learning needs. Generators learn the joint distribution of a real table and sample new records, but matching marginals and correlations does not guarantee that a classifier trained on synthetic data generalises to real data. The Train-Synthetic-Test-Real (TSTR) protocol [5] measures this utility directly.

A generator also assigns every synthetic record a label, and nothing guarantees that the label agrees with what the real data imply for that record. We study this with Synthetic Label Trustworthiness Estimation (SLTE), a pre-training check that scores each synthetic label against a model trained on real data, relabels the least consistent samples and down-weights uncertain ones. Our contributions are:

- a TSTR evaluation of five classifiers at six mixing ratios with ranking, threshold and calibration metrics and paired DeLong tests;
- SLTE, framed as an empirical label-consistency check, with a permutation null and a held-out real-label control that separate generator effects from properties of the correction rule;
- robustness checks across nine generation runs, three generators and two datasets;
- a real-only comparison at matched real-record counts showing that synthetic augmentation gave no net benefit here.

II. RELATED WORK

A. Synthetic Tabular Data Generation

Variational autoencoders [1] have been adapted to tabular data, where mixed column types need dedicated handling; CTGAN and its companion TVAE [2] introduced mode-specific normalisation, and CTGAN adds a conditional vector with training-by-sampling for imbalanced discrete columns. Gaussian copula models [3] and the diffusion-based TabDDPM [4] broaden the model families, and comparative TSTR studies report performance below real-data training [6].

B. Label Noise in Supervised Learning

Class-conditional label noise has been analysed [7] and corrected through an estimated noise transition matrix [8]. Cheng et al. [9] filtered instance-dependent noise with a confidence-regularised sample sieve, and Northcutt et al. [10] introduced confident learning, which flags labels that disagree with out-of-sample predicted probabilities. SLTE applies this idea to generated labels, using a model trained only on real data as the reference.

C. Synthetic Data in Healthcare Machine Learning

Adversarial networks have generated synthetic electronic health records [11], privacy-preserving synthetic health data have been evaluated [12], and a narrative review summarises applications and open problems [13]. Class imbalance is pervasive in clinical prediction, where accuracy can mislead [14]. Generated labels in imbalanced health data have received little direct study.

D. Tabular Machine Learning Benchmarks

Large benchmark studies report random forests among the strongest general-purpose classifiers [15] and tree-based ensembles outperforming deep learning on tabular data

Niyaz Ahmad Khan is an undergraduate researcher and student in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: ahmadkhanniyaz@gmail.com)
Md Yousuf Bin Nazib is an undergraduate student in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: yousufnazib888@gmail.com)
Eshrath Jahan Esha is an undergraduate student in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: eshajeshrat@gmail.com)
Golam Mostofa Naeem is a Senior Lecturer in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: gmnaeem7@gmail.com}
Islam Uddin is an undergraduate student in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: islamuddin3725@gmail.com)
Abu Ubayda is an undergraduate student in the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh. (e-mail: ubaydaazad@gmail.com)

[16], [17]. These benchmarks train on real data only.

III. METHODOLOGY

A. Datasets

The primary dataset is the Medical Appointment No-Show dataset [18]: 110,527 appointments at public clinics in Vitória, Brazil, with demographic, health and scheduling attributes and a binary no-show label (88,208 attended, 22,319 no-shows). To test generality we repeat the core analyses on the UCI Adult census dataset [19] (48,842 records; income above 50K for 23.9%), the benchmark on which CTGAN was introduced [2].

B. Preprocessing

The record with Age = −1 is removed. WaitDays (appointment minus scheduling date, clipped at zero) and the scheduling and appointment weekdays replace the two date columns, identifiers are dropped and the label is 1 for a no-show. The twelve predictors in Table I and 110,526 records are split into a fixed stratified 80/20 partition (88,420 training and 22,106 test records; random_state = 42). For Adult we drop the census weight and the education string (duplicated by education-num), frequency-encode the six nominal columns with training-split frequencies, keep "?" as a category and use the same split (39,073 training and 9,769 test records). Test records are never used for generation or correction.

TABLE I
FEATURE SCHEMA OF THE PRIMARY DATASET

Feature	Type	Encoding
Gender	Binary	M = 1, F = 0
Age	Integer	Years, 0 to 115
Neighbourhood	Nominal (81)	Frequency over all records
Scholarship, Hipertension, Diabetes, Alcoholism, SMS_received	Binary	0/1
Handcap	Ordinal	0 to 4
WaitDays	Integer	Days, clipped at 0
ScheduledDayOfWeek	Ordinal	0 (Monday) to 6
AppointmentDayOfWeek	Ordinal	0 (Monday) to 6
No-show (label)	Binary	Yes = 1, No = 0

Names follow the source file. Generation also used ScheduledMonth, dropped before evaluation.

C. Synthetic Data Generation

The primary synthetic set (run C0) comes from CTGAN (ctgan 0.12.1) with the library defaults: 300 epochs, batch size 500, generator and discriminator layers (256, 256), learning rates 2×10^{-4}, one discriminator step, pac = 10, embedding_dim = 128, generator_decay = discriminator_decay = 10^{-6} and log_frequency = True. It was trained on the twelve predictors plus ScheduledMonth and the label, with discrete_columns = [Gender, Scholarship, Hipertension, Diabetes, Alcoholism, Handcap, SMS_received, ScheduledDayOfWeek, AppointmentDayOfWeek, ScheduledMonth, No-show], and was not seeded. Its 88,420 samples are post-processed to valid ranges (Age 0 to 115, WaitDays ≥ 0, binary columns {0, 1}, Handcap {0, …, 4}).

With the same schema we add CTGAN runs with seeds 1 and 2 (C1, C2, via set_random_state), a matched run with log_frequency = False (N1, seed 1), TVAE [2] (T1; library defaults, no conditional vector) and a Gaussian copula with empirical marginals [3] (G1). For Adult we generate one CTGAN, TVAE and copula set (seed 1), declaring the nominal columns, sex, education-num and the label as discrete.

At sampling time, sample_original_condvec picks each row's conditioning column uniformly among the eleven discrete columns and a category from the stored probabilities, which log_frequency = True computes from log(count + 1). About one row in eleven is thus conditioned on the label and requests a no-show with probability 0.467 instead of 0.202, a direct effect that raises the minority share. The previously reported balances (0.1617 with, 0.2653 without) point the other way; Section V-I shows that the share varies more between identical runs (0.162 to 0.291) than between these two arms.

D. TSTR Experimental Protocol

Mixtures contain 0%, 20%, 40%, 60%, 80% or 100% synthetic records at a fixed training size (88,420; 39,073 for Adult), with real and synthetic rows sampled without replacement, and every model is tested on the real test partition. The classifiers are Logistic Regression (lbfgs, max_iter = 1000), Gaussian Naive Bayes, Decision Tree (max_depth = 10), Random Forest (100 trees) and Gradient Boosting (100 trees). Logistic Regression, Decision Tree and Random Forest use balanced class weights; Naive Bayes and Gradient Boosting have no class_weight parameter and are unweighted. SLTE weights enter every fit through sample_weight. On Adult, Logistic Regression uses standardised inputs because lbfgs does not converge on the unscaled capital-gain column.

Seed 42 is the reference; seeds 1 to 4 repeat the design with different mixture sampling and estimator seeds. Naive Bayes and Logistic Regression are deterministic given their data, Decision Tree and Gradient Boosting (no row or feature subsampling) are deterministic up to tie-breaking, and only Random Forest samples rows and features. At 0% and 100% every seed uses the same records in a different order, so seed variation there reflects estimator randomness only (Gradient Boosting at 100%: standard deviation below 0.0001).

AUC-ROC is the primary metric; we also report F1, precision and recall for the minority class, PR-AUC (average precision), balanced accuracy, macro-F1, the Brier score and the expected calibration error (ECE, ten bins). AUC differences are tested with the paired DeLong test [20] and Holm correction [21] within each family of comparisons. DeLong intervals reflect test-set sampling with the training data fixed, so they are conditional on one generation run; the separate runs assess variation between synthetic datasets.

IV. THE SLTE FRAMEWORK

Let (x^*, y^*) be a synthetic sample. Its label is corrupted if it disagrees with the real-data posterior:

$$y^* \neq \arg\max_y p_{\text{real}}(y \mid x^*) \quad (1)$$

The true posterior is unknown, so SLTE measures consistency with an estimate of it. A changed label therefore signals disagreement with a real-data model, not a verified error (Section V-E).

A. Stage 1: Label Confidence Score Computation

A Random Forest anchor (150 trees, balanced class weights) is trained on the standardised real training partition, and

$$\mathrm{LCS}_i = p_{\mathrm{anchor}}(y = y_i^* \mid x_i^*) \tag{2}$$

Across the 88,420 primary synthetic samples, LCS has mean 0.6977, standard deviation 0.2426 and skewness −0.9005, a left-skewed distribution concentrated near 1 (Fig. 1).

B. Stage 2: Topology-Aware KNN Label Correction

Samples with LCS < 0.40 (12,745) are relabelled by the majority vote of their seven nearest real training records (Euclidean distance on features standardised with training-partition statistics).

C. Stage 3: Uncertainty-Stratified Routing

Synthetic samples are weighted by LCS track (Table II), and real samples keep unit weight.

TABLE II
SLTE CONFIDENCE TRACK ASSIGNMENTS AND TRAINING WEIGHTS

Track	LCS Range	Count (%)	Weight
HIGH	≥ 0.70	52,682 (59.6%)	1.0
MID	0.40 ≤ LCS < 0.70	22,993 (26.0%)	0.7
LOW	< 0.40	12,745 (14.4%)	0.5

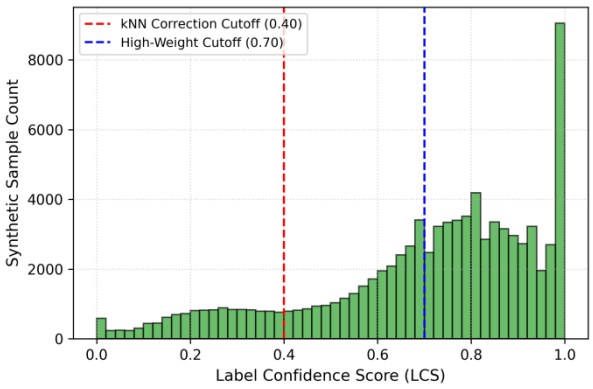

Fig. 1. Distribution of Label Confidence Scores across the synthetic dataset. Dashed lines mark the correction cutoff (0.40) and the high-confidence cutoff (0.70).

V. RESULTS AND ANALYSIS

A. Baseline Performance (100% Real Data)

With real training data only (0% columns of Tables III to V), AUC-ROC ranges from 0.6361 (Naive Bayes) to 0.7290 (Gradient Boosting). Accuracy misleads: Gradient Boosting reaches 79.80% accuracy but predicts only 32 positives among 22,106 test records at the default threshold (15 true, 17 false; recall 0.0034). Its average precision is 0.361 against a no-skill 0.202, a 1.79-fold lift (Fig. 2). Maximising F1 (t = 0.2195) gives precision 0.319 and recall 0.757; requiring recall ≥ 0.80 (t = 0.2015) gives precision 0.311.

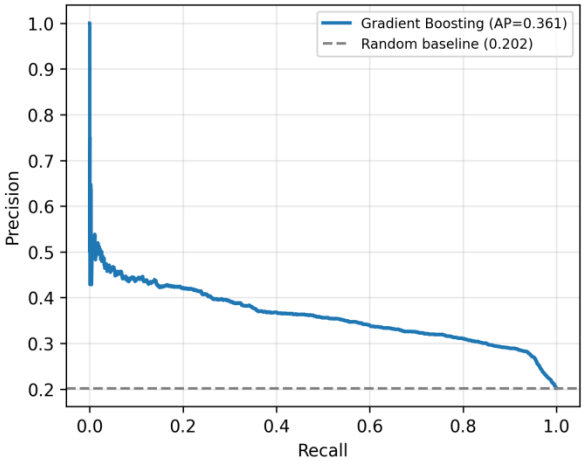

Fig. 2. Precision-recall curve for Gradient Boosting trained on 100% real data (AP = 0.361; random baseline = 0.202).

B. AUC-ROC Degradation Across Mixing Ratios

Without SLTE, AUC-ROC falls monotonically with the synthetic share for every classifier (Table III, Fig. 3). Across the five seeds the largest loss at 20% ranges from 0.96% to 1.11% (1.02% for seed 42, Logistic Regression). Against training on all real records the loss is significant in 22 of 25 DeLong comparisons, all except Decision Tree at 20% and 40% and Naive Bayes at 20%, so 20% is a low-degradation region rather than a safe threshold. Between 60% and 100%, Naive Bayes declines most in absolute terms (0.0535) and Gradient Boosting least (0.0254). At 100% the relative loss ranges from 5.72% (Gradient Boosting) to 11.46% (Naive Bayes).

TABLE III
AUC-ROC BY CLASSIFIER AND SYNTHETIC MIXING RATIO, WITHOUT AND WITH SLTE

Classifier	Condition	0%	20%	40%	60%	80%	100%
Decision Tree	No-SLTE	0.7157	0.7136	0.7104	0.6976	0.6878	0.6468
	SLTE	0.7157	0.7099	0.7060	0.6938	0.6759⁻	0.6499
Random Forest	No-SLTE	0.7106	0.7039	0.6986	0.6887	0.6724	0.6361
	SLTE	0.7106	0.7035	0.7007	0.6899	0.6839⁺	0.6851
Gradient Boosting	No-SLTE	0.7290	0.7251	0.7210	0.7127	0.7004	0.6873
	SLTE	0.7290	0.7232	0.7190	0.7137	0.7103⁺	0.7114
Logistic Regression	No-SLTE	0.6648	0.6580	0.6472	0.6339	0.6109	0.5889
	SLTE	0.6648	0.6543⁻	0.6432	0.6328	0.6209⁺	0.5962
Naive Bayes	No-SLTE	0.6361	0.6341	0.6284	0.6167	0.5918	0.5632
	SLTE	0.6361	0.6257⁻	0.6200⁻	0.6139	0.6081⁺	0.5987

Columns: synthetic share of training data. +/−: SLTE significantly higher/lower than no correction (paired DeLong test, Holm-adjusted p < 0.05, 25 tests).

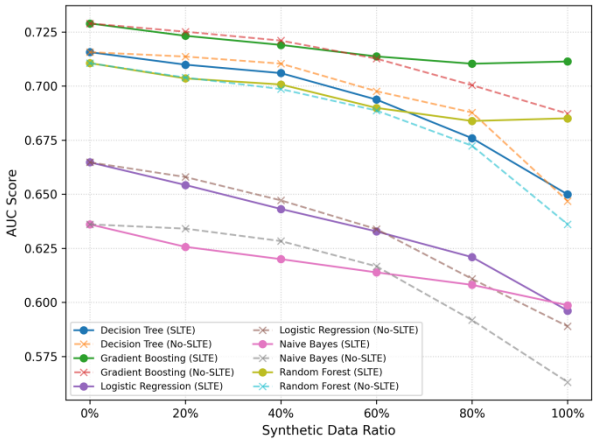

Fig. 3. AUC-ROC across synthetic mixing ratios for all five classifiers, with (solid) and without (dashed) SLTE correction.

C. F1 Score and Minority Class Collapse

F1 degrades more than AUC-ROC (Table IV, Fig. 4). From the real baseline to 100% synthetic, F1 falls by 8.8% for Decision Tree, the most stable classifier, 16.4% for Logistic Regression, 81.8% for Random Forest, 86.6% for Gradient Boosting and 90.5% for Naive Bayes, so stability does not follow the tree/parametric split. Gradient Boosting's F1 stays near zero because of its default threshold (Section V-A).

TABLE IV
MINORITY-CLASS F1 AT ALL SIX MIXING RATIOS, WITHOUT AND WITH SLTE

Classifier	Condition	0%	20%	40%	60%	80%	100%
Decision Tree	No-SLTE	0.4377	0.4370	0.4345	0.4328	0.4206	0.3993
	SLTE	0.4377	0.4348	0.4354	0.4274	0.4243	0.4043
Random Forest	No-SLTE	0.2629	0.2327	0.2105	0.1721	0.1236	0.0478
	SLTE	0.2629	0.2322	0.2050	0.1682	0.1030	0.0176
Gradient Boosting	No-SLTE	0.0067	0.0009	0.0004	0.0000	0.0000	0.0009
	SLTE	0.0067	0.0013	0.0031	0.0009	0.0027	0.0004
Logistic Regression	No-SLTE	0.4074	0.4091	0.3951	0.3781	0.3575	0.3407
	SLTE	0.4074	0.4095	0.3878	0.3766	0.3631	0.3029
Naive Bayes	No-SLTE	0.1872	0.1513	0.0953	0.0495	0.0253	0.0179
	SLTE	0.1872	0.2667	0.3155	0.3101	0.1892	0.0954

Default threshold of 0.50 for every classifier.

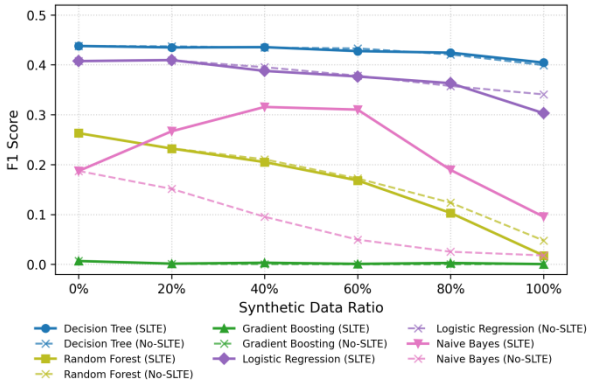

Fig. 4. Minority-class F1 across synthetic mixing ratios for all five classifiers, with (solid) and without (dashed) SLTE correction.

D. SLTE Framework Results

Of the 25 comparisons between SLTE and no correction (Table III), 11 are significant after Holm correction: seven gains, all at 80% or 100%, and four losses, three of them at 20% or 40%; none at 60% is significant. At 100%, Gradient Boosting improves by 0.0241 (95% CI [0.0185, 0.0297], z = 8.41, p < 0.001), recovering 57.8% of its loss (0.7114 against 0.6873 uncorrected and 0.7290 real). Section V-I shows how these effects vary between runs.

Threshold and calibration metrics qualify these gains (Table V). At 100% synthetic, SLTE raises PR-AUC for all five classifiers, but lowers recall for four of them (Logistic Regression from 0.562 to 0.355) and worsens both the Brier score and ECE for Random Forest, Gradient Boosting and Naive Bayes, while improving both for Decision Tree and Logistic Regression. Relabelling explains this: it moves the synthetic positive share from 16.2% to 4.8%, far below the real 20.2%, so corrected models rank better but predict fewer positives with shifted probabilities.

TABLE V
THRESHOLD AND CALIBRATION METRICS ON REAL AND FULLY SYNTHETIC TRAINING DATA

Classifier	Training	PR-AUC	Bal. acc.	Macro-F1	Recall	Brier	ECE
Decision Tree	Real	0.339	0.664	0.565	0.766	0.217	0.233
	Synthetic	0.292	0.622	0.555	0.644	0.221	0.231
	Syn. + SLTE	0.312	0.627	0.602	0.502	0.197	0.167
Random Forest	Real	0.350	0.558	0.563	0.203	0.159	0.080
	Synthetic	0.269	0.505	0.465	0.026	0.165	0.087
	Syn. + SLTE	0.323	0.503	0.453	0.009	0.179	0.158

Gradient Boosting	Real	0.361	0.501	0.447	0.003	0.145	0.011
	Synthetic	0.310	0.500	0.444	0.000	0.155	0.058
	Syn. + SLTE	0.350	0.500	0.444	0.000	0.181	0.165
Logistic Regression	Real	0.302	0.630	0.587	0.569	0.233	0.280
	Synthetic	0.260	0.561	0.506	0.562	0.243	0.289
	Syn. + SLTE	0.271	0.552	0.544	0.355	0.203	0.184
Naive Bayes	Real	0.290	0.533	0.528	0.129	0.167	0.054
	Synthetic	0.246	0.502	0.452	0.009	0.163	0.047
	Syn. + SLTE	0.278	0.517	0.489	0.054	0.170	0.108

Real: 0% synthetic. Synthetic: 100% synthetic. Brier and ECE: lower is better.

E. What the Label Changes Measure

Stage 2 changed 11,390 synthetic labels (12.88% of 88,420), 94.2% of them from no-show to attended (Table VI). First, 1,000 permutations of the synthetic labels, which keep the class balance but remove any feature-label association, give a 1→0 share of 94.35% (95% interval 94.20% to 94.49%); the observed share lies at the lower edge, so the asymmetry is no generator signature but the result of a seven-neighbour majority vote over reference data that are 80% majority class.

Second, we refitted the anchor and the neighbour model on 70,736 training records and applied SLTE to the genuine labels of the remaining stratified 17,684. It changed 2,511 of them (14.20%; denominator: all held-out records), 90.8% 1→0, overturning 63.8% of genuine no-shows and 1.6% of attendances. The same refitted anchor changes 12.81% of the synthetic labels (12.88% with the full anchor). SLTE thus changes genuine labels about as often as synthetic ones, so its change rate cannot estimate synthetic label corruption; it measures how far labels, real or synthetic, depart from a decision boundary that separates no-shows only moderately well (AUC-ROC 0.7290). At the synthetic positive share the held-out rates predict 11.70% changes; the excess of 1.11 points comes from synthetic positives being overturned more often (75.3% against 63.8%), the only sign that CTGAN's positive labels fit the real data less well.

TABLE VI
LABEL CHANGES MADE BY SLTE ON SYNTHETIC AND REAL LABELS

Labels	Anchor size	Changed (%)	1→0 share (%)	Positives overturned (%)	Negatives overturned (%)
Synthetic, run C0	88,420	12.88	94.2	75.0	0.89
Synthetic, run C0	70,736	12.81	95.0	75.3	0.76
Held-out real (17,684)	70,736	14.20	90.8	63.8	1.64

Overturned: share of each class changed.

F. Comparison Against Real-Only Training

Table VII trains each classifier on the real part of each mixture alone to test whether synthetic records add information. By point estimate, real-only training beat both mixed conditions in 17 of 20 classifier-ratio cells; the three exceptions are all Decision Tree: 20% (0.7102 against 0.7136 without and 0.7099 with SLTE), 40% (0.7058 against 0.7104 and 0.7060) and 80% (0.6663 against 0.6878 and 0.6759). With Holm correction over 40 tests, real-only training was significantly better in 15 of 20 cells against no correction and 15 of 20 against SLTE, and significantly worse only for Decision Tree at 80% without SLTE.

Gradient Boosting gives the strongest form of this result (Fig. 5): 17,684 real records alone reach 0.7273, within 0.0017 of the 0.7290 obtained from all 88,420 (DeLong p = 0.126), whereas adding 70,736 synthetic records to the same 17,684 lowers AUC-ROC to 0.7004 (0.7103 with SLTE). The synthetic records dilute the real signal rather than add to it.

TABLE VII
AUC-ROC: REAL-ONLY VERSUS MIXED TRAINING BY MIXING RATIO

Classifier	Condition	20%	40%	60%	80%
Decision Tree	Real-only	0.7102	0.7058	0.7005	0.6663
	Mixed	0.7136	0.7104	0.6976	0.6878†
	Mixed + SLTE	0.7099	0.7060	0.6938	0.6759
Random Forest	Real-only	0.7069	0.7075	0.7007	0.6934
	Mixed	0.7039	0.6986*	0.6887*	0.6724*
	Mixed + SLTE	0.7035	0.7007*	0.6899*	0.6839*
Gradient Boosting	Real-only	0.7294	0.7289	0.7287	0.7273
	Mixed	0.7251*	0.7210*	0.7127*	0.7004*
	Mixed + SLTE	0.7232*	0.7190*	0.7137*	0.7103*
Logistic Regression	Real-only	0.6653	0.6651	0.6636	0.6633
	Mixed	0.6580*	0.6472*	0.6339*	0.6109*
	Mixed + SLTE	0.6543*	0.6432*	0.6328*	0.6209*
Naive Bayes	Real-only	0.6384	0.6365	0.6329	0.6364
	Mixed	0.6341*	0.6284*	0.6167*	0.5918*
	Mixed + SLTE	0.6257*	0.6200*	0.6139*	0.6081*

*/†: real-only significantly better/worse (paired DeLong test, Holm-adjusted p < 0.05, 40 tests). Real records: 70,736, 53,052, 35,368 and 17,684.

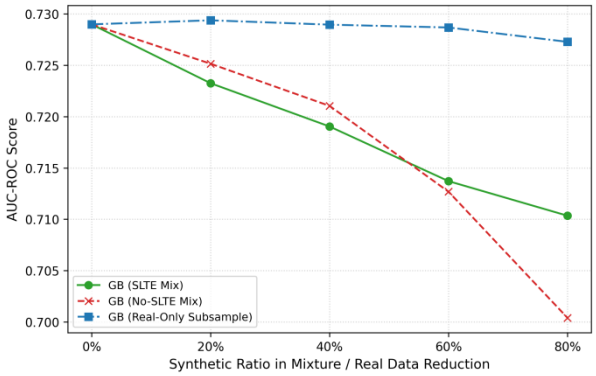

Fig. 5. Gradient Boosting AUC-ROC under three conditions: real-only training at matched real-record counts, uncorrected mixed training, and SLTE-corrected mixed training.

G. Component Ablation

At 100% synthetic (Table VIII), relabelling (Stage 2) provides most of the gain for Random Forest (0.0477 of 0.0490), Naive Bayes (0.0352 of 0.0355) and Gradient Boosting (0.0222 of 0.0241). For Decision Tree and Logistic Regression, weighting alone (Stage 3) gives higher AUC-ROC (0.6594 and 0.5990) than the full pipeline (0.6499 and 0.5962), so the two stages are not uniformly complementary.

TABLE VIII
COMPONENT ABLATION AT 100% SYNTHETIC SUBSTITUTION

Classifier	No-SLTE	Stage 2	Stage 3	Full
Decision Tree	0.6468	0.6481	0.6594	0.6499
Random Forest	0.6361	0.6837	0.6365	0.6851
Gradient Boosting	0.6873	0.7095	0.6978	0.7114
Logistic Regression	0.5889	0.5969	0.5990	0.5962
Naive Bayes	0.5632	0.5984	0.5786	0.5987

Stage 2: relabelling only. Stage 3: confidence weights only.

H. Reference-Set Size and Empirical Privacy

With 500, 1,000 or 5,000 randomly chosen reference records instead of 88,420 (Table IX), Gradient Boosting keeps 7%, 53% and 96% of its full-reference gain, but with 1,000 records SLTE falls below no correction for Decision Tree, Random Forest and Logistic Regression.

TABLE IX
REFERENCE-SET SIZE ABLATION AT 100% SYNTHETIC SUBSTITUTION

Classifier	No-SLTE	500	1,000	5,000	Full (88,420)
Decision Tree	0.6468	0.5490	0.6055	0.6232	0.6499
Random Forest	0.6361	0.6105	0.6144	0.6502	0.6851
Gradient Boosting	0.6873	0.6890	0.7000	0.7103	0.7114
Logistic Regression	0.5889	0.6043	0.5790	0.5963	0.5962
Naive Bayes	0.5632	0.5902	0.5743	0.5921	0.5987

The distance to closest record (DCR) [22], computed in standardised feature space on 5,000-record subsamples, screens for copying: the median synthetic-to-real distance is 0.246 against 0.068 between real records, and no sampled synthetic record coincides with a real one, whereas 1,353 sampled real records have an exact duplicate. DCR does not bound disclosure risk, which requires membership-inference evaluation [23], [24].

I. Robustness Across Runs, Generators and Datasets

Between CTGAN runs with identical settings the positive share ranged from 0.162 to 0.291 (C0, C1, C2; Table X); with the seed fixed, switching log_frequency off gave 0.224 (N1) against 0.291 (C1), lower, as the direct effect predicts, but inside that range. TVAE, which has no conditional vector, gave 0.177 and the copula 0.203.

Two results held in all nine runs. Full substitution lowered mean AUC-ROC (3.31% to 9.74% on the no-show data, 1.60% to 5.49% on Adult), with only Naive Bayes on T1, Logistic Regression on G1 and Naive Bayes on Adult CTGAN not losing, and training on the real part of each mixture alone was significantly better than mixed training in 6 to 30 of 40 comparisons and significantly worse in at most six.

On the no-show data SLTE changed synthetic labels only 1.11 to 1.91 points more often than held-out real labels would be at the same positive share, and the permutation null reproduced its 1→0 share to within one point. On Adult the excess reached 7.25 points for the copula and 3.01 points for CTGAN, and CTGAN's 1→0 share exceeded its null by 13.8 points.

SLTE's effect depended on run, classifier and dataset. On the no-show data its significant gains occurred mostly at 80% or 100% synthetic data, and at full substitution it improved Random Forest in all six runs and Gradient Boosting in four but lowered Decision Tree in three, Logistic Regression in three and Naive Bayes in two. On Adult it gave 7 to 18 significant gains and at most one significant loss per run.

TABLE X
SLTE AND AUC-ROC ACROSS GENERATION RUNS AND DATASETS

Run	Pos. share	Changed % (excess)	1→0 % (null)	Loss % (SLTE)	GB rec. %	SLTE +/−	Real better
No-show data (real positive share 0.202)							
C0, CTGAN	0.162	12.88 (+1.11)	94.2 (94.4)	9.74 (6.28)	57.8	7/4	30
C1, CTGAN	0.291	21.02 (+1.22)	97.6 (96.7)	4.45 (4.24)	45.5	4/9	22
C2, CTGAN	0.206	16.40 (+1.91)	96.6 (96.5)	8.12 (3.10)	78.6	10/1	25
N1, CTGAN*	0.224	17.15 (+1.71)	96.8 (96.3)	5.10 (4.81)	11.2	6/1	26

T1, TVAE	0.177	14.10 (+1.46)	94.4 (94.3)	3.31 (3.49)	46.3	3/9	13
G1, copula	0.203	16.06 (+1.76)	96.3 (96.2)	5.11 (5.99)	-25.2	5/10	21
Adult data (real positive share 0.239)							
CTGAN	0.303	13.22 (+3.01)	88.9 (75.1)	1.60 (1.10)	29.0	13/0	13
TVAE	0.226	10.13 (+1.18)	68.6 (64.5)	2.10 (1.33)	46.2	7/0	6
Copula	0.237	16.16 (+7.25)	78.3 (75.6)	5.49 (1.63)	61.4	18/1	19

*log_frequency = False. Excess: points above the held-out prediction. Null: permutation mean. Loss: mean relative AUC-ROC loss at 100% synthetic. GB rec.: Gradient Boosting loss recovered. Counts: significant DeLong tests (Holm) of 25 and 40.

VI. DISCUSSION

A. Practical Deployment Guidelines

Given these results, we recommend SLTE only when at least 80% of the training data must be synthetic, after checking on held-out real data that it helps the chosen classifier, and followed by threshold selection and recalibration because relabelling depletes the minority class. When real records exist, training on them alone is preferable to adding synthetic records from these generators. SLTE needs real labelled data, so it must run inside the data holder, which then releases only the relabelled, weighted synthetic set.

B. SLTE as a Label-Consistency Check

On the no-show data SLTE's changes are not detections of generator errors (Section V-E): permuted labels show the same asymmetry, and genuine labels are changed at a similar rate once the class mix is accounted for. SLTE is better described as an empirical label-consistency, or quality-control, step that aligns synthetic labels with the decision boundary learned from real data, as confident learning [10] does for annotated labels. The alignment improves ranking when synthetic data dominate training but imports the reference model's pull toward the majority class, which costs recall and, for some classifiers, calibration. It helped more on Adult, where the real-data model is stronger (Gradient Boosting AUC-ROC 0.92 against 0.73).

C. Limitations and Future Work

The evaluation covers two public datasets and binary outcomes. Generation variability rests on three CTGAN runs and one run of each other generator; the large differences between identical CTGAN runs mean that single-run TSTR results should be read with caution. The anchor and neighbour models use fixed hyperparameters, and the held-out check uses one split. Privacy evaluation is limited to DCR. The 20% and 80% boundaries are descriptive.

VII. CONCLUSION

On a medical no-show dataset and the Adult dataset, replacing real training records with synthetic ones lowered AUC-ROC in every generation run, and adding synthetic records to real ones rarely helped. SLTE recovered part of the loss when synthetic data dominated training but reduced recall and sometimes calibration, and its effect on some classifiers changed sign between runs. Synthetic data did not replace real data here; SLTE is useful only when substitution cannot be avoided and its benefit is confirmed on held-out real data.

STATEMENTS

Ethics: This study uses publicly available, de-identified datasets containing no direct patient identifiers. No new human-subject data were collected.

Data Availability: The no-show dataset is publicly available at https://www.kaggle.com/datasets/joniarroba/noshowappointments under its stated license, and the Adult dataset at https://doi.org/10.24432/C5XW20.

Code Availability: All preprocessing, generation and analysis code is available at https://github.com/theniyazkhan/slte.

Conflicts of Interest: The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Niyaz Ahmad Khan led the study from start to finish. He acquired the datasets, built the processing and modelling pipelines, ran the experiments and the SLTE analysis, wrote the first draft of the manuscript, prepared the revision, and handled all correspondence.

Md Yousuf Bin Nazib assisted in setting up the experimental framework, provided support during the data analysis phase, and critically reviewed the manuscript.

Eshrath Jahan Esha provided general support during the study and reviewed the manuscript.

Golam Mostofa Naeem contributed as a supervisor. He reviewed the manuscript to enhance its technical accuracy, clarity, coherence, and overall quality. Furthermore, he guided the research process and helped ensure the scientific rigor and academic integrity of the final manuscript.

Islam Uddin assisted with supporting tasks during the study and reviewed the manuscript.

Abu Ubayda provided limited support during the study and reviewed the manuscript.

All authors read and approved the final manuscript.

AI TOOL USAGE ACKNOWLEDGEMENT

Generative AI tools were used only in a supporting role while preparing this work. Claude and Gemini were utilized to help refine the research topic and structure the initial proposal. QuillBot was employed for language editing and tightening the wording of specific passages, while Perplexity was used to search for and verify relevant literature. The research design, data collection, analysis, results, and conclusions are entirely the authors' own. All AI-assisted text was thoroughly read and revised by the authors, who take full responsibility for the accuracy and originality of the work.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[2] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[3] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 2016, pp. 399–410.

[4] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "TabDDPM: Modelling tabular data with diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

[5] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," arXiv:1706.02633, 2017.

[6] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective table data synthesizing," in *Proc. Asian Conf. Mach. Learn. (ACML)*, 2021.

[7] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2013.

[8] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1944–1952.

[9] H. Cheng, Z. Zhu, X. Li, Y. Gong, X. Sun, and Y. Liu, "Learning with instance-dependent label noise: A sample sieve approach," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[10] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, 2021.

[11] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Proc. Mach. Learn. Healthcare Conf. (MLHC)*, 2017.

[12] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," *Neurocomputing*, vol. 416, pp. 244–255, 2020.

[13] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit. Health*, vol. 2, no. 1, 2023.

[14] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002.

[15] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[16] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022.

[17] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[18] J. Hoppen, "Medical appointment no shows," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/datasets/joniarroba/noshowappointments

[19] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996. doi: 10.24432/C5XW20.

[20] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[21] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[22] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, 2018.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2017, pp. 3–18.

[24] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data: Anonymisation groundhog day," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 1451–1468.

Niyaz Ahmad Khan is currently pursuing the B.Sc. degree in computer science and engineering at Metropolitan University, Sylhet, Bangladesh, with an expected graduation in 2026. His employment experience includes Engineer, Designer and Writer. He works with full-stack tools as well as relational and non-relational databases, to build data-driven applications. He proposed the SLTE framework for evaluating synthetic data quality in supervised machine learning, accepted at an international conference. Created an accessible, beginner-friendly Sylheti programming language designed to simplify computer science education. His research interests include machine learning, synthetic data trustworthiness, generative adversarial networks, low-resource natural language processing, handwriting recognition, and compiler design.

Md Yousuf Bin Nazib is currently pursuing the B.Sc. degree in computer science and engineering at Metropolitan University, Sylhet, Bangladesh, with an expected graduation in 2026. He has collaborated on several notable academic projects, including a course management system and an interactive forum platform. His research contribution includes serving as the second author on a paper presenting the SLTE framework for evaluating synthetic data quality in supervised machine learning, accepted at an international conference. His current research interests include machine learning, synthetic data evaluation, and test automation.

Eshrath Jahan Esha was born in Sylhet, Bangladesh. She is currently pursuing her degree at Metropolitan University, Sylhet, Bangladesh. She has a strong technical foundation with proficiency in multiple programming languages, including C, C++, Java, MySQL, and Python. Her practical experience includes the successful development of several projects utilizing Java and MySQL. In addition to her programming capabilities, she possesses foundational skills in digital design and has a strong passion for writing. Her current and future interests are heavily focused on conducting advanced academic and technological research.

Golam Mostofa Naeem is a Lecturer (Senior Grade) in the Department of Computer Science and Engineering at Metropolitan University, Bangladesh. His research interests include Machine Learning, Natural Language Processing (NLP), Explainable Artificial Intelligence (XAI), and Deep Learning. His research focuses on developing intelligent, interpretable, and reliable AI-driven approaches to address real-world challenges. In addition to his research activities, he is committed to advancing knowledge in artificial intelligence, exploring emerging computational techniques, and contributing to academic excellence through teaching, research, and scholarly collaboration.

Islam Uddin was born on May 27, 2001. He is expected to earn the B.Sc. degree in computer science and engineering from Metropolitan University, Sylhet, Bangladesh. His major field of study is computer science and engineering, with a strong emphasis on modern software development practices and scalable architectures. He is currently working as a Full Stack Developer and MERN Stack Specialist. His notable technical projects feature comprehensive client-server architectures. He is dedicated to continuous technological learning and is currently open to full-time professional software engineering opportunities.

Abu Ubayda was born in Sylhet, Bangladesh, in October 2002. He is currently pursuing the B.Sc. degree in computer science and engineering at Metropolitan University, Sylhet, Bangladesh. He has developed several academic software engineering projects, an e-commerce platform with a customized MySQL backend database for hardware retail management also engineered, a dynamic university management. He built an AI-integrated Telegram bot engineered to automate student routine tracking and faculty contact retrieval. His current research interests include artificial intelligence, cybersecurity, and the integration of computational tools with local history.