LEVERAGING GENERATIVE AI FOR ROBUST CLINICAL PREDICTIONS: A CASE STUDY ON DENTAL IMPLANT FAILURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning has the potential to improve clinical prediction and decision-making, yet its real-world effectiveness is often constrained by the characteristics of healthcare data. Clinical datasets are frequently small, heterogeneous, and highly imbalanced, limiting model generalization and reducing the detection of rare but clinically significant outcomes. These challenges motivate the development of learning frameworks that can effectively leverage limited data while preserving the statistical characteristics necessary for robust prediction. This paper investigates synthetic tabular data generation as a strategy for improving predictive modeling under data scarcity and class imbalance. We propose **GC-TabGAN** (Gaussian-Conditional Tabular GAN), a customized Generative Adversarial Network (GAN) for generating realistic synthetic clinical records, and integrate it with a Feed Forward Neural Network (**FFNNet Baseline**) for downstream prediction. Dental implant failure is used as a representative clinical case study because implant failures are relatively rare, arise from complex interactions among patient, clinical, and procedural factors, and therefore present a challenging prediction task. Using a public dataset of 747 implants, of which 117 (15.7%) failed, and a 10-fold cross-validation protocol in which the generator is refit within every training fold, we add 5%, 25%, 50%, and 90% of a 1,000-record fold-specific synthetic pool to the real training data. Synthetic augmentation mainly improves the detection of implant failures. Using only real data, the FFNNet Baseline achieved an accuracy of 0.92, an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.81, and a sensitivity to implant failure of 0.64. Augmenting the training data with synthetic samples raised the sensitivity to 0.76–0.80 while maintaining a specificity of 0.97–0.98; with 50% synthetic samples, the model reached an accuracy of 0.95, an AUC of 0.89. In contrast, oversampling the training data with SMOTE raised sensitivity to 0.88 but reduced specificity to 0.66, with an AUC of 0.77, well below that of GC-TabGAN, indicating that GC-TabGAN improves discrimination rather than merely shifting the decision boundary toward the minority class. These findings suggest that distribution-aware synthetic tabular data can partly offset data scarcity and class imbalance in small clinical datasets, a setting common in healthcare where large, balanced datasets are difficult to collect.

1 INTRODUCTION

Dental implants are a widely accepted solution for tooth replacement due to their durability, aesthetic appeal, and functional similarity to natural teeth. Although they demonstrate high overall success rates and patient satisfaction (Le Guéhennec et al., 2007; Duraccio et al., 2015; Moraschini et al., 2015), failure remains a persistent clinical concern: long-term reviews report cumulative survival of roughly 95% after ten or more years, so about one implant in twenty is lost (Moraschini et al., 2015; Howe et al., 2019). Beyond the obvious functional and psychological consequences for patients, implant failures also impose substantial clinical and economic burdens, often requiring additional surgeries, prolonged treatment times, and increased healthcare costs. These failures can arise from anatomical constraints, poor bone quality, surgical inaccuracies, systemic conditions such as osteoporosis and diabetes, or patient-related factors such as smoking (Levin, 2008; Raikar et al., 2017;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Vehemente et al., 2002). Clinical investigations have emphasized the importance of these variables: Gaêta-Araujo et al. (2020) identified inadequate spacing and anatomical perforations as major contributors; Camps-Font et al. (2018) highlighted postoperative infection and implant design; Naseri et al. (2020) confirmed a strong dose-dependent relationship between smoking and implant loss; and Annibali et al. (2016) observed higher risks of early failure among diabetic patients. Reviews further point to peri-implantitis, bone quality, and systemic disease as leading causes of implant complications (Gupta et al., 2023).

Given these multifactorial risks, researchers have increasingly turned to predictive modeling to support clinical decision-making. Traditional machine learning (ML) models, such as decision trees, logistic regression, and support vector machines, have shown moderate predictive ability but are often constrained by small and imbalanced datasets (Liu et al., 2018). More advanced approaches, including ensemble methods, and deep learning, have demonstrated superior performance, achieving accuracy and Area Under the Curve (AUC) values exceeding 0.90 in some cases (Gómez-Ríos et al., 2023; Şalgău et al., 2024; Wu et al., 2024; Shujaat, 2025; Ganz et al., 2021; Huang et al., 2022). However, issues of interpretability, generalizability, and data scarcity continue to limit their clinical translation. To overcome data limitations, recent studies have applied generative AI techniques such as Generative Adversarial Networks (GANs), Variational Auto Encoders (VAEs), and Wasserstein GANs (WGANs) to synthesize realistic datasets, improve robustness, and enable better model training in medical domains (Goodfellow et al., 2020; Arjovsky et al., 2017; Yi et al., 2019; Frid-Adar et al., 2018; Choi et al., 2017). These developments highlight the growing role of AI not only as a tool for data analysis but also as a transformative approach for personalized treatment planning and outcome prediction in dentistry.

Building on this body of work, the present study investigates whether synthetic tabular data can improve the prediction of dental implant failure when real clinical data are scarce and imbalanced. To this end, we propose GC-TabGAN, a tabular generator that handles features according to their distribution type, and evaluate it as a data augmentation method for a neural network classifier. Specifically, this study addresses the following research questions:

- **RQ1:** Under a leakage-free cross-validation protocol, does augmenting the real training data with synthetic records generated by GC-TabGAN improve implant failure prediction compared with training on real data alone?
- **RQ2:** How does predictive performance, and in particular the detection of implant failures (sensitivity), change as the proportion of synthetic data added to training increases?

The main contributions of this work are as follows:

- We propose GC-TabGAN, a lightweight generator for small, mixed-type clinical tables. It synthesizes approximately Gaussian features adversarially and samples categorical features from a Gaussian conditional model given the synthesized features, separately for each outcome class.
- We embed generation inside every cross-validation training fold, so no synthetic record depends on test data and all evaluation is performed on real records only.
- On a public dental implant dataset, we show that augmentation with GC-TabGAN raises sensitivity to implant failure from 0.64 to 0.76–0.80 without loss of specificity, whereas SMOTE oversampling gains sensitivity only at a large cost in specificity and AUC, and we analyze how performance and its fold-to-fold variability change with the synthetic fraction.

2 Related Work

Machine learning for implant failure prediction. Conventional machine learning methods have been applied to dental implant failure prediction using clinical and implant-related variables, establishing a foundation for data-driven risk assessment (Liu et al., 2018). In particular, Milić et al. (2026) compared logistic regression, random forest, and gradient boosting on the same publicly available dental implant dataset considered in this study, using stratified 5-fold cross-validation. Their findings demonstrate the potential of supervised learning for identifying implant failure patterns, but also reflect the limitations of relying exclusively on the size and class distribution of the original dataset.

Generative models for synthetic health data. Research has increasingly explored generative artificial intelligence for creating synthetic healthcare data. Ibrahim et al. (2025) reviewed GANs, variational autoencoders, diffusion models, and large language models across medical data modalities, emphasizing the need for clinically meaningful, privacy-aware, and standardized evaluation of synthetic data. For tabular data, a central difficulty is representing heterogeneous columns that mix categorical, discrete, and continuous variables. Widely used baselines include TableGAN (Park et al., 2018), CTGAN and TVAE (Xu et al., 2019), CTAB-GAN (Zhao et al., 2021), the diffusion-based TabDDPM (Kotelnikov et al., 2023), and CopulaGAN from the Synthetic Data Vault (Patki et al., 2016). For class imbalance specifically, interpolation-based oversampling such as SMOTE (Chawla et al., 2002) remains the standard, inexpensive alternative. Kang et al. (2025) introduced TT-GAN, which integrates transformer-based multi-head attention into a GAN to capture relationships among heterogeneous columns; its discretization and conversion stages allow it to represent complex, non-Gaussian distributions, and it outperformed CTGAN and CopulaGAN on the healthcare datasets evaluated. Bozkurt (2026) proposed sTableGAN, a stability-oriented modification of TableGAN that uses asymmetric discriminator updates, label smoothing, and Swish activations without explicit gradient penalties or feature matching; across several medical datasets it achieved lower distributional divergence, broader class coverage, competitive downstream utility, and limited evidence of memorization. In contrast to these general-purpose architectures, GC-TabGAN targets small clinical datasets with a lightweight design: approximately Gaussian features are synthesized adversarially, and binary and categorical features are sampled from a Gaussian conditional model given the synthesized features, with a separate model for each outcome class (Section 3.2).

Evaluating synthetic tabular data. As synthetic data generators become more sophisticated, several studies stress that they cannot be judged by a single similarity metric. Hernandez et al. (2025) proposed an evaluation framework covering fidelity, downstream utility, and privacy, and the trade-offs between them; they found that simpler models often achieved stronger fidelity and utility, while differential privacy improved protection at some cost to data quality. Kurakova & Homayouni (2025) introduced a four-dimensional framework covering quality, privacy, usability, and computational complexity, and proposed the Anomaly Proximity Score to detect abnormal or out-of-range synthetic records that conventional statistical measures may overlook.

Collectively, these studies indicate that GAN-based augmentation for dental implant data should preserve clinically relevant feature relationships, maintain minority-class representation, support downstream prediction, and be evaluated for both utility and privacy rather than for synthetic-to-real similarity alone. These gaps motivate the development and assessment of a generative framework tailored to dental implant data for improving predictive modeling under limited-data conditions.

3 METHODS

3.1 DATASET AND VARIABLES

We use the public dataset of Liu et al. (2018), which contains 747 records of individual dental implant procedures, 117 (15.7%) of which resulted in failure. The records include demographic, behavioral, and procedural variables, which are listed in Table 1.

Table 1: Clinical variables available in the dataset.

Cat.	Variables
Patient	Age, Gender, Systemic disease, Reason for tooth loss, Tobacco smoking, Betel nut chewing, Alcohol consumption
Clinical	Department, Surgeon experience, Location, Bone density, Timing, Ridge augmentation, Sinus augmentation, Implant system, Fixture length, Fixture width, Prosthesis type, Abutment angle, Fixation

Given a set of n data points $\left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{n}$, where $\mathbf{x}_i$ denotes the feature vector and y_i the corresponding label for the i-th data point, our goal is to learn a classification model to distinguish between dental implant success and failure. The original dataset codes implant failure as 0 and success as 1. Following the approach of Liu et al. (2018), we recode the target variable as $y_i \in 0, 1$,

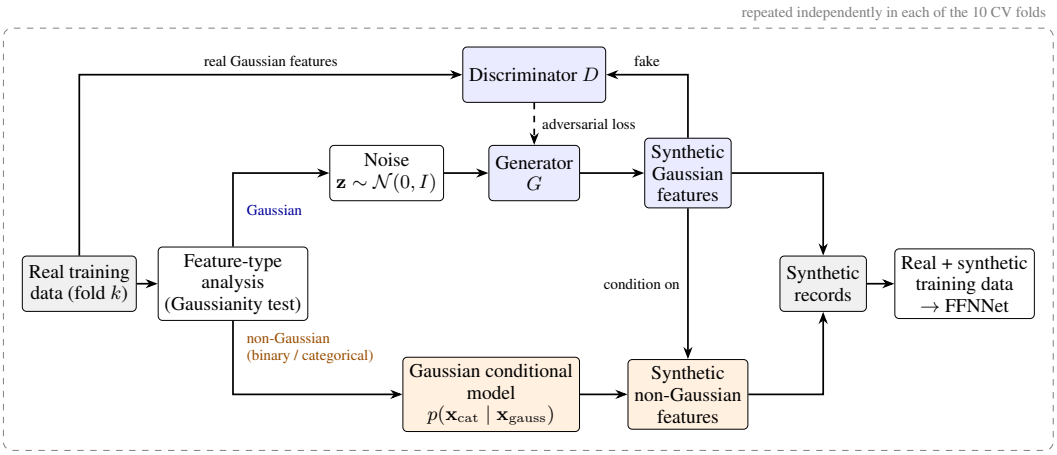

Figure 1: Overview of GC-TabGAN. Within each cross-validation fold, features of the real training data are split by distribution type. Gaussian features are synthesized by an adversarially trained generator–discriminator pair; non-Gaussian (binary/categorical) features are sampled from a Gaussian conditional model given the generated Gaussian features. The two parts are merged into complete synthetic records, which augment the real training data for the downstream FFNNet Baseline. Both parts are fit separately for each outcome class.

where $y_i = 1$ denotes dental implant failure and $y_i = 0$ denotes implant success (i.e., no failure). Thus, implant failure is treated as the positive class throughout this paper. Accordingly, sensitivity represents the proportion of implant failures correctly identified, whereas specificity represents the proportion of successful implants correctly identified. The feature vector $\mathbf{x}_i$ is constructed from the variables listed in Table 1. To ensure consistency and comparability with prior research, we adopt the methodology proposed by Liu et al. (2018), utilizing 10-fold cross-validation for model evaluation; with 747 records, each test fold contains roughly 75 implants, of which about 12 are failures. This evaluation strategy is uniformly applied throughout all subsequent modeling and experimental procedures.

To prevent information from the evaluation data leaking into the synthetic samples, data generation is performed entirely within the cross-validation loop. For each fold $k \in \{1, \ldots, 10\}$, the data are partitioned into a training set $\mathcal{D}_{\text{train}}^{(k)}$ (nine folds) and a held-out test set $\mathcal{D}_{\text{test}}^{(k)}$ (one fold). A new GC-TabGAN is then (i) trained from scratch using only $\mathcal{D}_{\text{train}}^{(k)}$; (ii) used to generate a pool of 1,000 synthetic records $\mathcal{S}^{(k)}$; and (iii) a fraction of $\mathcal{S}^{(k)}$ is appended to $\mathcal{D}_{\text{train}}^{(k)}$ to train the downstream classifier, which is finally evaluated on $\mathcal{D}_{\text{test}}^{(k)}$. The held-out fold is therefore never seen by either the generator or the classifier, and all test sets consist exclusively of real records. Likewise, standardization, normality testing, and discretization (Section 3.2) are fit on $\mathcal{D}_{\text{train}}^{(k)}$ only.

We developed a four-layer Feed Forward Neural Network (FFNNet), referred to as the FFNNet Baseline, to address the dental implant failure prediction task. This architecture consists of one input layer, two hidden layers with ReLU activations , and one sigmoid output layer. The full model details, including the hyperparameter search strategy, are provided in Appendix A.2.

To test the hypothesis that incorporating synthetic data can enhance predictive performance, we train the same FFNNet Baseline on real data augmented with records generated by GC-TabGAN; in Table 2, this configuration is denoted "+ GC-TabGAN". As a reference for conventional imbalance handling, we also train the same FFNNet Baseline, with the same hyperparameters and decision threshold, on training data oversampled with SMOTE (Chawla et al., 2002), which creates synthetic failures by interpolating between neighboring real failures. SMOTE is applied to $\mathcal{D}_{\text{train}}^{(k)}$ only, so the test folds remain unchanged real records; this configuration is denoted "FFNNet Baseline + SMOTE". SMOTE is implemented with the `imbalanced-learn` library (Lemaître et al., 2017) (`imblearn.over_sampling`) using random seed 10.

3.2 GC-TabGAN: Model Details

GC-TabGAN extends the standard GAN framework with a Gaussian conditional component (overview in Figure 1). Continuous features that are approximately Gaussian are synthesized adversarially, while binary and categorical features are sampled from a conditional model given the synthesized Gaussian features. All components described below are fit using only the training data $\mathcal{D}_{\text{train}}^{(k)}$ of the current fold, and separately for each outcome class (see *Class-conditional generation* below).

Notation and feature typing. Each record is written as $\mathbf{x} = (\mathbf{x}^g, \mathbf{x}^c)$, where $\mathbf{x}^g \in \mathbb{R}^{d_g}$ collects the features modeled as Gaussian and $\mathbf{x}^c = (c_1, \ldots, c_{d_c})$ collects the non-Gaussian features, with $c_j \in \{1, \ldots, K_j\}$. Binary and categorical variables (e.g., gender, smoking, implant system) are assigned to $\mathbf{x}^c$ by definition. Each continuous variable (e.g., age, fixture length, fixture width) is tested for normality with the Shapiro–Wilk test on the training fold; it is assigned to $\mathbf{x}^g$ if the null hypothesis of normality is not rejected at level $\alpha = 0.05$, and is then standardized using the training-fold mean and standard deviation. Continuous variables that fail the test are discretized into quantile-based bins and subsequently treated as categorical variables.

Adversarial model for Gaussian features. The generator $G_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_g}$ maps a white-noise vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$ to a synthetic vector $\tilde{\mathbf{x}}^g = G_\theta(\mathbf{z})$ in the standardized feature space. The discriminator $D_\phi : \mathbb{R}^{d_g} \to (0, 1)$ outputs the probability that its input is a real record. Both networks are multilayer perceptrons; their layer sizes and activations are listed in Appendix A.2. To stabilize training on the small dataset, zero-mean white noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ (instance noise) is added to both real and generated inputs of D_ϕ; σ is listed in Appendix A.1. The networks are trained with the standard adversarial objective (Goodfellow et al., 2020):

$$\mathcal{L}_D(\phi) = -\,\mathbb{E}_{\mathbf{x}^g \sim \mathcal{D}_{\text{train}}^{(k)}}\big[\log D_\phi(\mathbf{x}^g)\big] \;-\; \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big[\log\big(1 - D_\phi(G_\theta(\mathbf{z}))\big)\big], \tag{1}$$

$$\mathcal{L}_G(\theta) = -\,\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big[\log D_\phi(G_\theta(\mathbf{z}))\big], \tag{2}$$

where Eq. equation 2 is the non-saturating generator loss; both losses are binary cross-entropies. D_ϕ and G_θ are updated alternately, with n_D discriminator steps per generator step.

Gaussian conditional model for non-Gaussian features. For each non-Gaussian feature c_j, we model the Gaussian features conditionally on each of its levels,

$$p(\mathbf{x}^g \mid c_j = l) = \mathcal{N}\big(\mathbf{x}^g;\, \boldsymbol{\mu}_{jl},\, \boldsymbol{\Sigma}_{jl}\big), \qquad p(c_j = l) = \pi_{jl}, \qquad l = 1, \ldots, K_j. \tag{3}$$

The parameters are fit by maximum likelihood on the training fold: π_{jl} is the relative frequency of level l, $\boldsymbol{\mu}_{jl}$ is the mean of $\mathbf{x}^g$ over records with $c_j = l$, and $\boldsymbol{\Sigma}_{jl}$ is their sample covariance, regularized as $\boldsymbol{\Sigma}_{jl} + \lambda\mathbf{I}$ with $\lambda = 10^{-6}$ to remain well conditioned for rare levels. Given a synthetic Gaussian vector $\tilde{\mathbf{x}}^g$, each non-Gaussian feature is then sampled from its posterior,

$$p(c_j = l \mid \tilde{\mathbf{x}}^g) = \frac{\pi_{jl}\,\mathcal{N}(\tilde{\mathbf{x}}^g; \boldsymbol{\mu}_{jl}, \boldsymbol{\Sigma}_{jl})}{\sum_{m=1}^{K_j} \pi_{jm}\,\mathcal{N}(\tilde{\mathbf{x}}^g; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})}, \qquad \tilde{c}_j \sim \text{Categorical}\big(p(c_j = \cdot \mid \tilde{\mathbf{x}}^g)\big). \tag{4}$$

Non-Gaussian features are sampled independently of one another given $\tilde{\mathbf{x}}^g$.

Class-conditional generation. Generation is class-conditional. Within each fold, a separate generator–discriminator pair and a separate conditional model (Eqs. equation 3–equation 4) are fit to the training records of each outcome class, and every synthetic record inherits the label of the model that produced it.

Sampling procedure. After training in fold k, a synthetic record of class y is produced by (i) drawing $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$ and computing $\tilde{\mathbf{x}}^g = G_{\theta,y}(\mathbf{z})$; (ii) sampling $\tilde{c}_1, \ldots, \tilde{c}_{d_c}$ from Eq. equation 4 using the conditional model of class y; and (iii) mapping $\tilde{\mathbf{x}}^g$ back to the original units by inverting the standardization. Repeating this for both classes until 1,000 records are obtained gives the synthetic pool $\mathcal{S}^{(k)}$ described in Section 3.1.

Training details. Both networks are trained with Adam ($\eta_G = \eta_D = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) for $E = 10$ full-batch epochs, each consisting of one discriminator update followed by one generator update. The FFNNet Baseline is trained with binary cross-entropy and Adam ($\eta_C = 10^{-4}$) for $E_C = 20$ epochs with a batch size of $B_C = 8$, and predicts failure when its output exceeds 0.5. To control randomness and ensure reproducibility, fixed random seeds are used for the cross-validation split, network initialization, and noise sampling; reported results are based on a single fixed seed. The full set of hyperparameters and the software environment are listed in Appendices A.1 and A.2.

4 Results and Discussion

Table 2 compares the performance of several machine learning models, including our FFNNet Baseline trained on real data only and trained on real data augmented with GC-TabGAN records, following the experimental protocol established by Liu et al. (2018). All values are means (and standard deviations, for our models) over the 10 cross-validation folds. The results for the first nine models are taken directly from Liu et al. (2018). Accuracy should be compared with care. Only 15.7% of implants failed, so a classifier that always predicts success already reaches an accuracy of 0.84. Moreover, the accuracies reported by Liu et al. (2018) equal the average of their sensitivity and specificity (e.g., Decision Tree: $(0.59 + 0.77)/2 = 0.68$), which suggests that they were computed on class-balanced data. We therefore base comparisons with prior work mainly on sensitivity, specificity, and AUC. For GC-TabGAN, the number in parentheses is the fraction of the 1,000 fold-specific synthetic samples that was added to the real training data (Section 3.1). Among the traditional models reported by Liu et al. (2018), Bagging combined with Decision Trees achieved the highest performance, with an accuracy of 0.70 and an AUC of 0.74. Despite its strong specificity (0.82), its sensitivity was relatively low at 0.58; that is, it identified only 58% of implant failures. None of the previously reported models exceeded a sensitivity of 0.61. Our FFNNet Baseline achieved an accuracy of 0.92, a sensitivity of 0.64, a specificity of 0.98, and an AUC of 0.81. The average of its mean sensitivity and specificity is 0.81, compared with at most 0.70 for the models of Liu et al. (2018). Its sensitivity is higher than that of all previously reported models, but it still missed roughly one in three implant failures, while almost all successful implants were classified correctly.

Table 2 also includes the more recent results of Milić et al. (2026), who evaluated L1-regularized logistic regression, Random Forest, and Gradient Boosting on the same dataset. These results are not strictly comparable to ours because they were obtained under stratified 5-fold rather than 10-fold cross-validation, and they are reported without standard deviations. With all features, Random Forest was their best model, with an accuracy of 0.86 and an AUC of 0.79, followed by Gradient Boosting (0.85; 0.76) and logistic regression (0.73; 0.73). Both ensembles identified successful implants with high specificity (0.96–0.97) but did not report failure-class sensitivity. The FFNNet Baseline exceeds these models in accuracy (0.92) and AUC (0.81), and GC-TabGAN widens the margin further (up to 0.95 and 0.89).

With GC-TabGAN augmentation, performance improved at all augmentation levels, and the largest gain was in the detection of implant failures: sensitivity increased from 0.64 for the FFNNet Baseline to 0.76 at the 5% and 25% levels, 0.79 at 50%, and 0.80 at 90%, while specificity remained at 0.97–0.98. The 50% configuration gave the highest accuracy (0.95) and AUC (0.89, tied with 90%). The 90% configuration achieved the highest sensitivity among the GC-TabGAN configurations and the lowest fold-to-fold variability (s.d. 0.08 versus 0.13–0.16), suggesting that larger amounts of synthetic data may also stabilize failure detection. At this level, the 900 synthetic records outnumber the roughly 672 real training records of each fold. The differences between augmentation levels (e.g., 0.76 versus 0.80 sensitivity) are small relative to the fold-to-fold standard deviations and should not be over-interpreted.

Comparison with SMOTE. Oversampling with SMOTE produced a markedly different trade-off. It achieved the highest sensitivity in Table 2 (0.88), but its specificity fell from 0.98 to 0.66, and its accuracy (0.70) and AUC (0.77) were lower than those of the FFNNet Baseline trained on real data alone (0.92 and 0.81). The average of sensitivity and specificity was 0.77 for SMOTE, compared with 0.81 for the FFNNet Baseline and 0.89 for GC-TabGAN at the 50% level. At the class ratio of

Table 2: Classification performance under 10-fold cross-validation (mean ± s.d.). Implant failure is the positive class (sensitivity = proportion of failures detected). Rows 1–9 are reported by Liu et al. (2018); their accuracy appears to have been computed on class-balanced data (see text). Rows 10–12 are reported by Milić et al. (2026) using all features under stratified *5-fold* cross-validation. The standard deviations were not reported in either study. For "+ GC-TabGAN" rows, the FFNNet Baseline is trained on real data plus the fraction (in parentheses) of the 1,000 synthetic samples. For "FFNNet Baseline + SMOTE", the training folds are oversampled with SMOTE. Best value in each column is shown in bold; – = not reported.

Classifier	Accuracy	Sensitivity	Specificity	AUC
DT	0.68	0.59	0.77	0.67
LGR	0.62	0.61	0.64	0.64
SVM	0.63	0.58	0.68	0.63
Bagging + DT	0.70	0.58	0.82	0.74
Bagging + LGR	0.63	0.57	0.68	0.67
Bagging + SVM	0.62	0.60	0.64	0.67
AdaBoost + DT	0.67	0.47	0.87	0.74
AdaBoost + LGR	0.60	0.51	0.69	0.66
AdaBoost + SVM	0.63	0.51	0.75	0.65
LR (L1)[‡]	0.73	–	0.76	0.73
Random Forest[‡]	0.86	–	0.96	0.79
Gradient Boosting[‡]	0.85	–	0.97	0.76
FFNNet Baseline	0.92±0.04	0.64±0.14	**0.98**±0.02	0.81±0.07
FFNNet Baseline + SMOTE	0.70±0.05	**0.88**±0.09	0.66±0.05	0.77±0.05
+ GC-TabGAN (0.05)	0.94±0.03	0.76±0.13	**0.98**±0.02	0.87±0.07
+ GC-TabGAN (0.25)	0.94±0.04	0.76±0.16	**0.98**±0.01	0.87±0.08
+ GC-TabGAN (0.50)	**0.95**±0.03	0.79±0.14	**0.98**±0.02	**0.89**±0.07
+ GC-TabGAN (0.90)	0.94±0.01	0.80±0.08	0.97±0.01	**0.89**±0.04

[‡]Milić et al. (2026) label the class with 630 records as "failure"; in the original coding of the dataset this is the successful-implant class. Their reported recall is therefore the proportion of successful implants correctly identified and is listed here as specificity. Sensitivity to implant failure was not reported. Accuracy and AUC do not depend on the choice of positive class.

the dataset, these mean rates imply that SMOTE would miss about 2 failures per 100 implants but raise about 29 false alarms, so that only about one in three predicted failures would be a true failure. GC-TabGAN at the 50% level would miss about 3 failures and raise about 2 false alarms per 100 implants, with roughly nine in ten predicted failures being correct. These figures are derived from mean sensitivity and specificity and are approximate.

The comparison also helps to separate the effect of class rebalancing from that of the synthetic records themselves. Both methods increase the share of failures in the training data, but to different degrees. The GC-TabGAN pool is balanced (500 records per class), so adding 50% or 90% of it raises the share of failures from about 16% to about 30% and 35%, respectively, whereas SMOTE, with its default setting, balances the classes fully. Part of SMOTE's lower specificity at the fixed 0.5 threshold may therefore reflect this stronger shift in class balance, which threshold tuning could partly recover. AUC, however, does not depend on the class balance of the training data or on the decision threshold. Here GC-TabGAN reached 0.87–0.89, compared with 0.77 for SMOTE and 0.81 for the FFNNet Baseline trained on real data alone; the difference between SMOTE and the real-data baseline is within fold-to-fold variation. This suggests that the records generated by GC-TabGAN add information that helps the classifier separate the two classes, whereas SMOTE's interpolations between neighboring failures mainly move the decision boundary. A plausible reason is that linear interpolation is ill-suited to the many categorical variables in this dataset (Table 1), whereas GC-TabGAN samples them from a conditional model.

Figure 2 compares the performance of our baseline model and the proposed GC-TabGAN using ROC curves. Unfortunately, the models of Liu et al. (2018) could not be included due to the unavailability of results at a sufficiently granular level. Across augmentation levels, the gains in sensitivity and AUC are largely reached by the 50% level, and adding more synthetic data yields little further improvement on average.

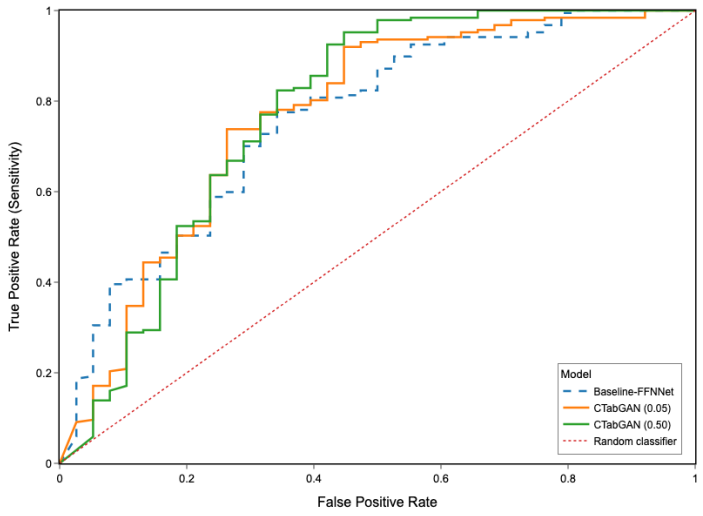

Figure 2: ROC curves of the FFNNet Baseline trained on real data only and on real data augmented with 50 and 500 GC-TabGAN records (5% and 50% of the synthetic pool). Both augmented models achieve a higher AUC than the baseline.

5 Limitations and Future Work

Despite the promising outcomes of this study, several limitations must be acknowledged:

- **Limited Dataset and Generalization:** The dataset comprises only 747 records from a homogeneous source. This limits the model's generalizability to broader populations or different clinical settings. Differences in demographics, habits, or treatment protocols across regions could impact performance (Huang et al., 2022). Future studies should include diverse, multi-institutional datasets to validate robustness.
- **Optimal Ratio of Synthetic Data:** Four synthetic data injection levels (5%, 25%, 50%, and 90%) were evaluated, with 50% yielding the best performance. However, the ideal synthetic-to-real data ratio remains unclear and may vary by dataset and task. Further studies should explore finer-grained levels and adaptive mixing strategies.
- **Comparison with Other Augmentation Methods:** The comparison with SMOTE suggests that the gains of GC-TabGAN are not explained by class rebalancing alone, but it covers a single resampling method at a fixed decision threshold. Comparisons with class weighting, random oversampling, threshold-tuned SMOTE, and other tabular generators such as CTGAN, TVAE (Xu et al., 2019), CTAB-GAN (Zhao et al., 2021), and TabDDPM (Kotelnikov et al., 2023), together with paired significance tests (Demšar, 2006), are needed to fully isolate the contribution of GC-TabGAN.
- **Single Dataset, Seed, and Threshold:** Results are based on one public dataset, a single random seed, and a fixed decision threshold of 0.5. With about 12 failures per test fold, fold-level sensitivity is noisy. Repeated cross-validation, threshold-independent metrics such as the area under the precision–recall curve and balanced accuracy, and evaluation on additional small clinical tabular datasets would strengthen the conclusions.
- **Limited Evaluation of Synthetic Data Quality:** The study assessed synthetic data quality indirectly via model performance metrics (accuracy, AUC, etc.). Future evaluations could incorporate distributional similarity analysis, feature importance comparisons, privacy measures such as the distance to the closest real record (Hernandez et al., 2025; Kurakova & Homayouni, 2025), or expert review to better assess the realism of GAN-generated data (Choi et al., 2017).

- **Model Interpretability and Clinical Adoption:** While deep learning models like the FFNNet Baseline and GC-TabGAN provide strong predictive accuracy, their lack of interpretability poses challenges for clinical adoption. Incorporating explainability tools such as SHAP or LIME may improve trust and facilitate clinical use (Doshi-Velez & Kim, 2017).

Before clinical deployment, it must also be ensured that synthetic data do not introduce bias or obscure patient-specific nuances (Wiens et al., 2019; Gianfrancesco et al., 2018). Such work should follow established reporting guidelines, namely TRIPOD+AI for prediction models (Collins et al., 2024) and, for prospective evaluation, CONSORT-AI and SPIRIT-AI (Liu et al., 2020; Cruz Rivera et al., 2020), and should involve both patients and clinicians in shared decision-making (Shortliffe & Sepúlveda, 2018; Sambasivan et al., 2021; Sperling et al., 2025).

6 Conclusion

This study demonstrates the potential of integrating GANs and deep learning to enhance predictive modeling in healthcare, specifically in the context of dental implant failure prediction. Our FFNNet Baseline outperformed previously reported traditional machine learning approaches such as Decision Trees, Support Vector Machines, and ensemble methods in AUC (0.81 versus at most 0.74), sensitivity to implant failure (0.64 versus at most 0.61), and the average of sensitivity and specificity (0.81 versus at most 0.70); it also exceeded the AUC (0.79) of the best Random Forest model recently reported by Milić et al. (2026), although their results were obtained under 5-fold cross-validation. Augmenting the training data with GC-TabGAN further increased sensitivity to 0.76–0.80 without loss of specificity, with the best accuracy and AUC observed at the 50% level. By contrast, SMOTE oversampling raised sensitivity to 0.88 only by lowering specificity to 0.66, and its AUC (0.77) remained well below that of GC-TabGAN, which indicates that the benefit of GC-TabGAN goes beyond rebalancing the classes.

These findings are consistent with the growing evidence that synthetic data can help address the limitations of small, imbalanced clinical datasets (Choi et al., 2017; Ibrahim et al., 2025). Additionally, performance was stable across synthetic-data ratios, suggesting that the method is not overly sensitive to this choice. However, challenges remain: GC-TabGAN still misses about one in five implant failures, and interpretability and generalizability across diverse populations remain open questions. Addressing them, as outlined in Section 5, could allow synthetic augmentation to support earlier identification of high-risk implant cases.

AI Use Statement

Generative AI tools were used to assist with language editing and software development. The authors designed the methodology, conducted the experiments, verified all generated code and text, analyzed the results, and take full responsibility for the manuscript.

References

Susanna Annibali, Nicola Pranno, Maria Paola Cristalli, Gerardo La Monaca, and Antonella Polimeni. Survival analysis of implant in patients with diabetes mellitus: a systematic review. *Implant dentistry*, 25(5):663–674, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Mustafa Hakan Bozkurt. Bridging stability and utility in synthetic tabular data: The sTableGAN model. *Knowledge-Based Systems*, 336:115303, 2026. doi: 10.1016/j.knosys.2026.115303.

Octavi Camps-Font, Pablo Martín-Fatás, Adrià Clé-Ovejero, Rui Figueiredo, Cosme Gay-Escoda, and Eduard Valmaseda-Castellón. Postoperative infections after dental implant placement: Variables associated with increased risk of failure. *Journal of periodontology*, 89(10):1165–1173, 2018.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pp. 286–305. PMLR, 2017.

Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, Maarten van Smeden, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, and SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, 26(9): 1351–1363, 2020. doi: 10.1038/s41591-020-1037-7.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Donatella Duraccio, Federico Mussano, and Maria Giulia Faga. Biomaterials for dental implants: current and future trends. *Journal of Materials Science*, 50:4779–4812, 2015.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Hugo Gaêta-Araujo, Nicolly Oliveira-Santos, Arthur Xavier Maseti Mancini, Matheus Lima Oliveira, and Christiano Oliveira-Santos. Retrospective assessment of dental implant-related perforations of relevant anatomical structures and inadequate spacing between implants/teeth using cone-beam computed tomography. *Clinical oral investigations*, 24:3281–3288, 2020.

Nancy Ganz, Alicia E. Ares, and Horacio D. Kuna. Procedure to improve the accuracy of dental implant failures by data science techniques. *Journal of Computer Science and Technology*, 21(2): e13, 10 2021. doi: 10.24215/16666038.21.e13.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018.

Inmaculada Gómez-Ríos, Esteban Egea-López, and Antonio José Ortiz Ruiz. ORIENTATE: automated machine learning classifiers for oral health prediction and research. *BMC Oral Health*, 23 (1):408, 2023. doi: 10.1186/s12903-023-03112-w.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Anirudh Gupta, Bhairavi Kale, Deepika Masurkar, and Priyanka Jaiswal. Etiology of dental implant complication and failure—an overview. *AIMS Bioengineering*, 10(2):141–152, 2023. doi: 10.3934/bioeng.2023010.

Mikel Hernandez, Pablo A. Osorio-Marulanda, Mikel Catalina, Lorea Loinaz, Gorka Epelde, and Naiara Aginako. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Frontiers in Digital Health*, 7:1576290, 2025. doi: 10.3389/fdgth.2025.1576290.

Maria-Sophia Howe, William Keys, and Derek Richards. Long-term (10-year) dental implant survival: A systematic review and sensitivity meta-analysis. *Journal of Dentistry*, 84:9–21, 2019. doi: 10.1016/j.jdent.2019.03.008.

Nannan Huang, Peng Liu, Youlong Yan, Ling Xu, Yuanding Huang, Gang Fu, Yiqing Lan, Sheng Yang, Jinlin Song, and Yuzhou Li. Predicting the risk of dental implant loss using deep learning. *Journal of Clinical Periodontology*, 49(9):872–883, 2022.

Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gökhan Ertaylan, and Michel Dumontier. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, 189:109834, 2025. doi: 10.1016/j.compbiomed.2025.109834.

Ha Ye Jin Kang, Minsam Ko, and Kwang Sun Ryu. Tabular transformer generative adversarial network for heterogeneous distribution in healthcare. *Scientific Reports*, 15(1):10254, 2025. doi: 10.1038/s41598-025-93077-3.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17564–17579, 2023.

Anastasia Kurakova and Hajar Homayouni. A comprehensive evaluation framework for synthetic medical tabular data generation. *Journal of Biomedical Informatics*, 171:104939, 2025. doi: 10.1016/j.jbi.2025.104939.

Laurent Le Guéhennec, Assem Soueidan, Pierre Layrolle, and Yves Amouriq. Surface treatments of titanium dental implants for rapid osseointegration. *Dental materials*, 23(7):844–854, 2007.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

Liran Levin. Dealing with dental implant failures. *Journal of applied oral science*, 16:171–175, 2008.

Chia-Hui Liu, Cheng-Jyun Lin, Ya-Han Hu, and Zi-Hung You. Predicting the failure of dental implants using supervised learning techniques. *Applied Sciences*, 8(5):698, 2018. doi: 10.3390/app8050698.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara Darzi, Jonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.

Marija S. Milić, Vladimir S. Todorović, Milan Vučetić, and Nataša Milosavljević. Predicting dental implant failure using machine learning: comparative evaluation of Random Forest, gradient boosting, and logistic regression with feature importance analysis. *Computer Methods in Biomechanics and Biomedical Engineering*, 2026. doi: 10.1080/10255842.2026.2645167. Published online 18 March 2026.

V. Moraschini, L. A. da C. Poubel, V. F. Ferreira, and E. dos S. P. Barboza. Evaluation of survival and success rates of dental implants reported in longitudinal studies with a follow-up period of at least 10 years: a systematic review. *International journal of oral and maxillofacial surgery*, 44 (3):377–388, 2015. doi: 10.1016/j.ijom.2014.10.023.

Roohollah Naseri, Jaber Yaghini, and Awat Feizi. Levels of smoking and dental implants failure: a systematic review and meta-analysis. *Journal of Clinical Periodontology*, 47(4):518–528, 2020.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018. doi: 10.14778/3231751.3231757.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, 2016. doi: 10.1109/DSAA.2016.49.

Sonal Raikar, Pratim Talukdar, Sarala Kumari, Sangram Kumar Panda, Vinni Mary Oommen, and Arvind Prasad. Factors affecting the survival rate of dental implants: A retrospective study. *Journal of International Society of Preventive and Community Dentistry*, 7(6):351–355, 2017.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21):2199–2200, 2018.

Sohaib Shujaat. Automated machine learning in dentistry: A narrative review of applications, challenges, and future directions. *Diagnostics*, 15(3), 2025. ISSN 2075-4418. doi: 10.3390/diagnostics15030273. URL https://www.mdpi.com/2075-4418/15/3/273.

Jessica Sperling, Whitney Welsh, Erin Haseley, Stella Quenstedt, Perusi B Muhigaba, Adrian Brown, Patti Ephraim, Tariq Shafi, Michael Waitzkin, David Casarett, and Benjamin A Goldstein. Machine learning-based prediction models in medical decision-making in kidney disease: patient, caregiver, and clinician perspectives on trust and appropriate use. *Journal of the American Medical Informatics Association*, 32(1):51–62, 2025. doi: 10.1093/jamia/ocae255.

Valerie A Vehemente, Sung-Kiang Chuang, Shadi Daher, Ali Muftu, and Thomas B Dodson. Risk factors affecting dental implant survival. *Journal of Oral Implantology*, 28(2):74–81, 2002.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Ziang Wu, Xinbo Yu, Feng Wang, and Chun Xu. Application of artificial intelligence in dental implant prognosis: A scoping review. *Journal of Dentistry*, pp. 104924, 2024.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective table data synthesizing. In *Proceedings of the 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 97–112, 2021.

Cristiana Adina Șalgău, Anca Morar, Andrei Daniel Zgarta, Diana-Larisa Ancuța, Alexandros Rădulescu, Ioan Liviu Mitrea, and Andrei Ovidiu Tănase. Applications of machine learning in periodontology and implantology: A comprehensive review. *Annals of biomedical engineering*, 52(9):2348–2371, 2024. doi: 10.1007/s10439-024-03559-0.

A Appendix

A.1 Synthetic Data Generation: Implementation Summary

For each outcome class, the generator G_θ maps $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$ through two ReLU layers (15 and 30 units) to a linear output. The discriminator D_ϕ uses two ReLU layers (25 and 50 units) followed by a sigmoid output, and receives inputs perturbed by instance noise of standard deviation σ. Both are trained with binary cross-entropy using full-batch updates and one discriminator step per generator step. Exact values are given in Table 3.

Table 3: Hyperparameters of GC-TabGAN and the FFNNet Baseline, using identical configurations across all folds.

Component	Hyperparameter	Value
Feature typing	Normality test / level α	Shapiro–Wilk / 0.05
Generator G_θ	Noise dimension d_z; distribution	d; $\mathcal{N}(0, I)$
	Hidden layers (units, activation)	15 ReLU; 30 ReLU
	Output layer (units, activation)	d, Linear
Discriminator D_ϕ	Hidden layers (units, activation)	25 ReLU; 50 ReLU
	Output layer (units, activation)	1, Sigmoid
	Instance-noise std. σ	0.05
GAN training	Loss	Binary cross-entropy
	Optimizer; learning rates (G / D)	Adam; 10^{-3} / 10^{-3}
	Adam (β_1, β_2)	$(0.9,\ 0.999)$
	Weight initialization (first layer)	He uniform
	Epochs; batch size; n_D	10; full batch $(N_c)^\dagger$; 1
Conditional model	Covariance regularization λ	10^{-6}
Synthetic pool	Records per fold (failure / success)	$1,000$ (500 / 500)
FFNNet Baseline	Optimizer	Adam
	Batch size; early stopping	10; yes
SMOTE baseline	Implementation; random seed	`imblearn.over_sampling`; 10
General	Software; hardware	PyTorch / NVIDIA H100 GPU

†One generator is trained per class c on its N_c training records; each epoch performs a single discriminator update on $2N_c$ samples (N_c real, N_c generated) followed by a single generator update on N_c noise vectors.

Table 4: Hyperparameter optimization configuration for the FFNNet baseline.

HPO Parameter	Value
HPO framework	Ray Tune
Search strategy	Random sampling
Number of configurations	50
Optimization objective	Validation loss
Objective direction	Minimize
Scheduler	ASHA
Maximum epochs per trial	100
Grace period	1 epoch
Reduction factor	2
Hidden layer 1 (h_1)	$\{16, 32, 64, 128, 256\}$
Hidden layer 2 (h_2)	$\{16, 32, 64, 128, 256\}$
Learning rate	Log-uniform, 10^{-4}–10^{-1}

A.2 Hyperparameter Optimization

Hyperparameters of the FFNNet Baseline were optimized using Ray Tune with an Asynchronous Successive Halving Algorithm (ASHA) scheduler. The optimization objective was validation loss, computed on a validation split drawn from the training data, which was minimized across 50 sampled configurations. The search space included the widths of the two hidden layers and the learning rate. Specifically, the hidden-layer widths h_1 and h_2 were selected from $\{16, 32, 64, 128, 256\}$, and the learning rate was sampled from a log-uniform distribution over $[10^{-4}, 10^{-1}]$. Each trial was allowed to run for a maximum of 100 epochs. ASHA used a grace period of one epoch and a reduction factor of two to terminate underperforming configurations early. The selected configuration is reported in Table 4.