Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Conditional Diffusion-Based Data Augmentation for Imbalanced Multi-Mode Fault Diagnosis in Predictive Maintenance

ABDELHAFID HAMDI ALAOUI

ab.hamdialaoui@gmail.com

Université Hassan II Casablanca ENSAM: Universite Hassan II Casablanca Ecole Nationale Superieure d'Arts et Metiers https://orcid.org/0000-0001-6513-1130

ANWAR MEDDAOUI

Research Article

Keywords: Predictive maintenance, Industry 4.0,class imbalance, data augmentation, diffusion models, generative models, fault diagnosis

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10208619/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Conditional Diffusion-Based Data Augmentation for Imbalanced Multi-Mode Fault Diagnosis in Predictive Maintenance

Dr. ABDELHAFID HAMDI ALAOUI[1*]
ENSAM, Hassan II university, Casablanca, Morocco
ab.hamdialaoui@gmail.com
ORCID: 0000-0001-6513-1130

Prof. ANWAR MEDDAOUI[2]
ENSAM, Hassan II university, Casablanca, Morocco
anwar.meddaoui@ensam-casa.ma
ORCID: 0000-0001-9308-1679

*: Corresponding author

ABSTRACT

Predictive maintenance (PdM) in Industry 4.0 relies on data-driven models that learn the signatures of impending failures from condition-monitoring data. In real operating conditions, however, fault events are scarce and unevenly distributed across failure modes, producing severe class imbalance that biases learning toward the healthy class and degrades the detection of exactly the rare, high-cost failures that PdM is meant to anticipate. This paper investigates conditional generative data augmentation as a remedy, and in particular evaluates a class-conditional denoising diffusion probabilistic model (DDPM) for synthesizing minority-class samples. Using the publicly available AI4I 2020 Predictive Maintenance dataset, we cast the problem as a five-class, single-label fault-diagnosis task with four rare failure modes and benchmark the proposed diffusion augmentation against no augmentation, random oversampling, SMOTE, a conditional GAN (cGAN) and a conditional VAE (cVAE). Each augmentation strategy is assessed both intrinsically, through distributional fidelity metrics (maximum mean discrepancy and a classifier two-sample test), and extrinsically, through downstream diagnosis performance measured with two independent classifiers (a multilayer perceptron and a random forest) over eight random seeds. The diffusion-based augmentation yields the highest rare-failure recall and the highest balanced accuracy with both classifiers, improving minority-class recall from 0.61 to 0.76 (multilayer perceptron) and from 0.41 to 0.76 (random forest), with the improvement over every baseline being consistent across all eight runs (Wilcoxon signed-rank test, $p = 0.008$). A transparent precision recall trade-off is observed and analyzed. The results indicate that diffusion models are a competitive and reproducible tool for mitigating class imbalance in predictive maintenance, especially when the operational priority is to maximize the detection of rare failures.

Keywords: Predictive maintenance, Industry 4.0,class imbalance, data augmentation, diffusion models, generative models, fault diagnosis.

I Introduction

The fourth industrial revolution has equipped manufacturing assets with dense networks of sensors that continuously stream temperature, rotational-speed, torque and wear measurements. Predictive maintenance (PdM) exploits these streams to anticipate equipment failure and schedule interventions before breakdowns occur, reducing unplanned downtime, extending asset life and lowering maintenance cost [1, 2]. Machine-learning and deep-learning models are now central to this paradigm, and recent systematic reviews map the rapid growth of AI-driven predictive maintenance in Industry 4.0 and the open challenges that remain [3, 4]. Yet their effectiveness is bounded by a structural property of industrial data this translates into the rarity of failures and breakdowns. A healthy machine generates the overwhelming majority of observations, while each distinct failure mode is represented by only a handful of examples [5].

This imbalance is not a nuisance but the core difficulty of PdM. Standard classifiers trained under empirical risk minimization implicitly optimize overall accuracy, which can be made almost perfect by always predicting the majority healthy class [5]. The cost of this behavior is borne precisely where it matters most the rare failure modes, whose timely detection is the entire purpose of a PdM system, are systematically missed. Re-weighting losses and resampling the training set are common countermeasures [5, 6], but classical resampling either duplicates existing minority points (random oversampling), adding no new information, or interpolates between them (SMOTE) [7], which can blur class boundaries and fabricate implausible samples in regions the true data never occupy.

Deep generative models offer an attractive alternative rather than copying or interpolating, they learn the conditional distribution of each failure mode and can synthesize novel yet realistic samples. Generative adversarial networks (GANs) [8, 9] and variational autoencoders (VAEs) [10]have been explored for this purpose, but GAN training is notoriously unstable on small tabular datasets and VAEs tend to produce over-smoothed samples. Denoising diffusion probabilistic models (DDPMs)[11, 12] have recently set a new standard in generative modeling through a stable, likelihood-based training objective, yet their use for tabular condition-monitoring data in predictive maintenance remains comparatively underexplored [13].

This paper addresses that gap with a controlled, reproducible benchmark. We formulate fault diagnosis on the AI4I 2020 dataset [14] as an imbalanced five-class problem, implement a class-conditional diffusion model for minority-class augmentation, and compare it head-to-head against five baselines under an identical protocol. The contributions are the following. First, we provide a uniform experimental framework in which random oversampling, SMOTE, a conditional GAN, a conditional VAE and a conditional diffusion model are evaluated with the same target balance, the same downstream classifiers and the same untouched real test set. Second, we evaluate each method on two complementary axes distributional fidelity of the synthetic data and downstream diagnostic utility rather than on a single score. Third, we report results across eight random seeds with non-parametric significance testing, and we analyze transparently the precision recall trade-off that diffusion augmentation induces. All code, hyper-parameters and random seeds are described in sufficient detail to reproduce the study.

II Related Work

II.1 Predictive maintenance and the imbalance problem

Data-driven predictive maintenance has grown from classical condition-monitoring and reliability analysis into a mature branch of Industry 4.0, and several recent systematic reviews chart its dominant models, methods and open challenges[1, 3, 4, 15]. A theme that recurs across these surveys is that aggregate accuracy is a poor proxy for operational value because of the rarity of failures a model that simply predicts the healthy state can appear excellent while missing precisely the events maintenance is meant to anticipate [5, 16]. Closely related to the present study, imbalance-aware and explainable predictive-maintenance frameworks have recently been evaluated on the same AI4I 2020 benchmark [20], which confirms both the relevance of this dataset and the practical weight of the imbalance question. Our work differs from such studies in that it isolates the data-augmentation dimension and benchmarks it systematically, rather than proposing a single end-to-end pipeline.

II.2 Resampling and classical augmentation

The most widely adopted countermeasure for imbalance is resampling at the data level[13, 17] Random oversampling duplicates minority instances, it is trivial to apply but adds no new information and he can even promote overfitting to the few repeated points. The Synthetic Minority Over-sampling Technique (SMOTE) [9] improves on this by interpolating between a minority sample and its nearest minority neighbors, and it remains a strong baseline across domains. Its core assumption that the minority manifold is locally linear is, however, fragile when the minority classes are tiny and multimodal, as the failure modes of an industrial machine typically are interpolating between two distinct failure mechanisms can place synthetic points in low-density or class-overlapping regions that the real process never visits. This limitation is the conceptual starting point for generative augmentation.

II.3 Generative augmentation with GANs and VAEs

Generative models reframe augmentation as the task of learning, and then sampling from, the conditional distribution of each class. Conditional generative adversarial networks [8, 18] and their tabular variants [4] produce class-conditioned samples by training a generator against a discriminator, while conditional variational autoencoders [10, 19] offer a stable, likelihood-based decoder with an explicit latent prior. Both families have been applied to industrial and fault-diagnosis data, with mixed outcomes: GANs are expressive but notoriously unstable to train on the small tabular datasets typical of PdM, where mode collapse can silently discard entire failure modes, whereas VAEs train reliably but tend to produce over-smoothed, low-diversity samples that add little beyond what interpolation already provides.

II.4 Diffusion models and tabular data

Denoising diffusion probabilistic models[11, 12] generate data by learning to reverse a gradual noising process. They combine a stable, likelihood-based objective with high sample quality, and have rapidly become the de-facto standard in image generation. Their extension to heterogeneous tabular data most prominently Tab-DDPM[13] has shown that the same recipe transfers to mixed numerical and categorical features. Early applications to imbalanced fault diagnosis have recently begun to appear, mostly on vibration signals[20, 21]. Despite this momentum, diffusion-based augmentation for imbalanced condition-monitoring data in

predictive maintenance, evaluated rigorously against both classical and generative baselines on fidelity and downstream utility and on a single controlled tabular benchmark, remains comparatively underexplored. Filling that gap, under a single controlled protocol and with non-parametric significance testing, is the contribution of this paper.

III Materials & Methods

Our experimental design follows a single guiding principle: to isolate the contribution of the augmentation strategy itself, every other component of the pipeline is held fixed and every safeguard against optimistically biased estimates is enforced. The AI4I 2020 records are first reduced to a clean, single-label five-class problem and projected into a common eight-dimensional feature space; a single stratified partition then separates a training set from a real test set that is never altered thereafter. Augmentation whether a classical resampler or a deep generative model acts exclusively on the training partition, rebalancing the four minority failure modes to a common target while leaving the test distribution exactly as observed. Each augmented training set feeds two structurally different classifiers, whose predictions on the untouched real data are examined along two complementary axes: the distributional fidelity of the synthetic samples and their downstream diagnostic utility. To prevent conclusions that hold only for a fortunate split or initialization, the whole procedure is repeated across eight random seeds and the resulting paired differences are assessed with a non-parametric significance test. Figure 1 renders this pipeline end to end, and the subsections that follow formalize each stage

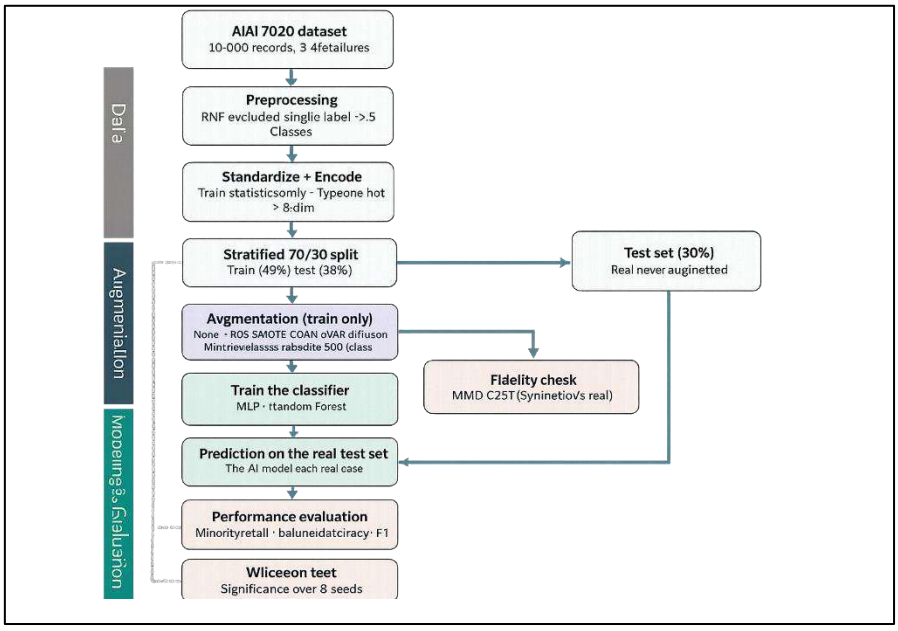


Figure 1. A flow chart of the feature selection heuristic

III.1 Dataset

We use the AI4I 2020 Predictive Maintenance dataset[22], a publicly available benchmark of 10,000 records simulating an industrial milling machine. Each record contains a product quality variant (Type ∈ {L, M, H}) and five process measurements air temperature, process

temperature, rotational speed, torque and tool wear together with a binary machine-failure label and five independent failure-mode flags: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF) and random failure (RNF).

The overall failure rate is 3.4%, producing a severe imbalance that closely mirrors real PdM settings.

Because RNF is by construction, a random event uncorrelated with the process variables, it is unpredictable from the features and is excluded, following common practice. We further restrict the data to single-label records by removing the small number of rows in which more than one real failure mode is active simultaneously, which yields a clean five-class target: a majority |No Failure| class and the four minority failure modes. The resulting class counts are reported in Table 1 and visualized in Figure 2, the four failure modes together represent less than 4% of the samples, with the rarest mode (TWF) accounting for only a few dozen records.

It is worth recalling how each mode is triggered in the simulator [1], because this shapes how hard each is to learn.

Tool wear failure (TWF) occurs at a randomly drawn tool-wear time, so it is only weakly tied to the instantaneous features and is the hardest to predict.

Heat dissipation failure (HDF) is driven by a small temperature difference combined with low rotational speed.

Power failure (PWF) depends on the product of torque and rotational speed lying outside an operating band.

Overstrain failure (OSF) is governed by the product of tool wear and torque exceeding a Type-dependent threshold. Two consequences follow for modeling. First, the machine Type is genuinely informative it sets the OSF threshold which justifies keeping it as a feature. Second, the failure modes occupy different, partly overlapping regions of the feature space, so a useful augmenter must respect this multi-modal structure rather than smear the minority classes together.

Table 1. Class distribution of the single-label AI4I 2020 subset used in this study.

Class	Description	Count	Share (%)
No Failure	Healthy operating point	9670	96.92
TWF	Tool wear failure	43	0.43
HDF	Heat dissipation failure	106	1.06
PWF	Power failure	80	0.80
OSF	Overstrain failure	78	0.78

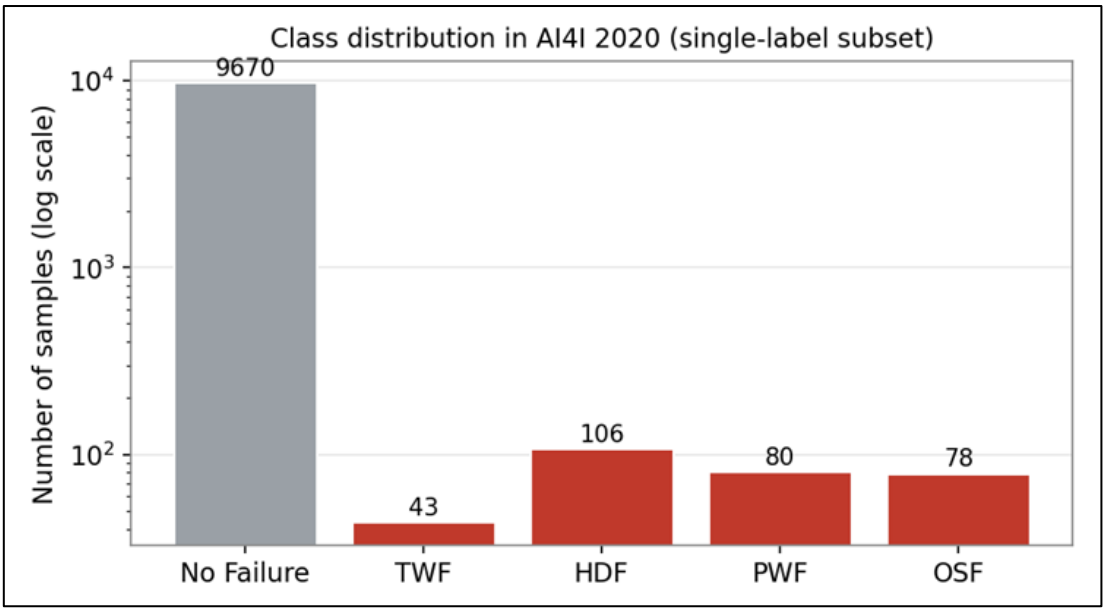


Figure 2. Class distribution

III.2 Problem formulation

Let a sample be described by a feature vector $x \in \mathbb{R}$ obtained by standardizing the five numeric process variables(Air-Temperature, Process-Temperature, Rotational-Speed, Torque, Tool Wear) and appending a three-dimensional one-hot encoding of the machine Type(L,M,H), and let $y \in \{0,1,2,3,4\}$ denote the class, where 0 is the healthy class and 1to 4 index the failure modes TWF, HDF, PWF and OSF. The task is to learn a classifier $f: x \rightarrow y$ that maximizes the detection of the minority classes under severe imbalance. Data augmentation acts only on the training partition: for each minority class c, a generator synthesizes additional samples so that the training count of class c reaches a common target N_target, leaving the test partition entirely real and untouched.

III.3 Preprocessing and Augmentation Target

Numeric features are standardized with statistics estimated on the training split only, preventing information leakage from the test split. The categorical Type is one-hot encoded. For every augmentation method the four minority classes are raised to N_target = 500 training samples each, while the majority class is left unchanged; this fixed, method-independent target isolates the effect of the generator from any difference in the amount of augmentation. For the generative models the synthesized continuous features are kept as produced, and the generated Type dimensions are projected back to a valid one-hot vector by an arg-max operation, ensuring that all methods operate in the same eight-dimensional feature space.

III.4 Augmentation Methods

We compare five augmentation strategies against a no-augmentation baseline, ranging from naive replication to modern generative modeling. The two classical baselines come from the imbalanced-learn toolkit [17]: random oversampling (ROS), which simply draws minority instances with replacement, and SMOTE [7], which creates new points by interpolating between a minority sample and its nearest minority neighbors. Neither invents genuinely new

structure ROS repeats what is already there, and SMOTE stays on the line segments between existing points which is exactly the limitation that motivates a generative approach.
We then add two well-established deep generators. The conditional GAN (cGAN)[8, 18] trains a generator against a discriminator, both told which class they are dealing with the generator turns a Gaussian latent vector and a one-hot class code into a synthetic feature vector, following the tabular-GAN conditioning of Xu et al. [9]. The conditional VAE (cVAE)[10, 19] takes a different route, learning a class-conditioned encoder decoder with a Gaussian prior and producing new samples by decoding latent draws. So that the comparison reflects the generator family rather than how much data each model is shown, the three conditional generators (cGAN, cVAE and the diffusion model described next) are all trained jointly on the pooled minority data, which lets the four failure modes share statistical strength a real advantage when the rarest mode has only a few dozen examples.
Our proposed augmenter is a class-conditional denoising diffusion probabilistic model (DDPM)[11, 12], applied directly to the standardized tabular features in the spirit of recent tabular-diffusion work [13]. A forward Markov process gradually corrupts a real sample x_0 into Gaussian noise over T steps according to a fixed variance schedule

$$\beta_1, \dots, \beta_T, with\ \alpha_t = 1 - \beta_t\ and\ \bar{\alpha}_t = \prod_{s \le t} \alpha_s$$

The closed-form marginal of the forward process is :

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t;\ \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right), x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\,\varepsilon, \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \qquad (1)$$

A neural network $\varepsilon_\theta(x_t, t, c)$, conditioned on the diffusion step t and the class c, is trained to predict the injected noise by minimizing the simplified denoising objective

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,\varepsilon,t,c}\ \|\varepsilon - \varepsilon_\theta\ (\sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\ \varepsilon, t, c)\|\text{^}2 \qquad (2)$$

New minority samples are then drawn by ancestral sampling from pure noise x_T ~ $\mathcal{N}$(0, I), iterating the learned reverse transition

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\ (x_t, t, c)\right) + \sigma_t\ z, \qquad z \sim \mathcal{N}(0, I)\ \ (z = 0\ \text{à}\ t = 1) \qquad (3)$$

The denoiser is a multilayer perceptron with sinusoidal-free learned step embeddings and a one-hot class condition, trained for 6,000 iterations with T = 200 diffusion steps and a linear β schedule from 1×10^{-4} to 2×10^{-2}. The cGAN and cVAE use multilayer-perceptron architectures of comparable capacity and are trained for 4,000 iterations, ensuring a fair comparison of generator families rather than of training budgets.

III.5 Downstream Classifiers

To verify that conclusions are not an artifact of a single classifier, two structurally different downstream models are trained on each augmented training set and evaluated on the same real test set: a multilayer perceptron (MLP) with two hidden layers (64 and 32 units) and a random forest (RF) of 300 trees. Neither classifier uses class-weighting, so any improvement is attributable to the augmentation rather than to cost-sensitive learning.

III.6 Evaluation Protocol and Metrics

The data are split into a 70% training and 30% stratified test partition. Augmentation and classifier training are performed on the training partition only; all metrics are computed on the real test partition. The whole pipeline is repeated over eight random seeds that jointly control the split, the generator initialization and the classifier initialization, and we report the mean and standard deviation across seeds.

Distributional fidelity of the synthetic minority data is quantified with two intrinsic measures: the maximum mean discrepancy (MMD) with a radial-basis-function kernel (lower is better) and a classifier two-sample test (C2ST), in which a random forest attempts to distinguish real from synthetic samples and the cross-validated area under the ROC curve is reported (a value near 0.5 indicates that synthetic and real samples are statistically indistinguishable). Downstream utility is measured with macro-averaged F1, balanced accuracy, and, focusing on the rare classes, the macro-averaged recall, precision and F1 over the four failure modes. Because in PdM a missed failure is typically far costlier than a false alarm, minority recall and balanced accuracy are treated as the primary operational metrics. Statistical significance of the difference between the proposed diffusion method and each baseline is assessed with the two-sided Wilcoxon signed-rank test [23]over the eight paired seeds.

III.7 Implementation Details and Hyper-parameters

All generators and classifiers are deliberately small, both because the minority training data are scarce a few hundred samples in total and because we want the comparison to reflect the modeling principle rather than raw capacity. The three deep generators therefore share multilayer-perceptron backbones of comparable width, and the two downstream classifiers are left at sensible defaults without any class-weighting. Table 2 collects the full configuration, the values were fixed once, before looking at test performance, and were not tuned per method, so that no baseline is handicapped by a smaller training budget. The conditional generators are trained jointly on the pooled minority data, and at sampling time the three one-hot Type dimensions are projected back to a valid category by an arg-max, keeping every method in the same eight-dimensional feature space.

Table 2. Hyper-parameter configuration of the generators

Component	Setting	Value
cGAN	generator/discriminator MLP width	128
cGAN	latent dimension	16
cGAN	optimizer / learning rate	Adam / 2×10^{-2}
cGAN	iterations, batch size	4000, 64
cVAE	encoder/decoder MLP width	128
cVAE	latent dimension	12
cVAE	optimizer / learning rate	Adam / 1×10^{-3}
cVAE	KL weight β, iterations	0.5, 4000
Diffusion	denoiser MLP (3 hidden layers)	128
Diffusion	diffusion steps T	200
Diffusion	β schedule (linear)	$1\times10^{-4} \rightarrow 2\times10^{-2}$
Diffusion	optimizer / learning rate	Adam / 1×10^{-3}
Diffusion	iterations, batch size	6000, 64
MLP classifier	hidden layers	(64, 32)
MLP classifier	optimizer, max iterations	Adam, 400
Random forest	number of trees	300
Protocol	augmentation target N_target	500 / class
Protocol	train/test split, seeds	70/30 stratified, 8

IV Results

IV.1 Distributional Fidelity of the Synthetic Data

Table 3 reports the intrinsic fidelity of the synthetic minority samples. SMOTE attains the lowest MMD and a C2ST AUC closest to 0.5, which is expected because interpolation reproduces the local geometry of the real data almost exactly, this apparent advantage, however, reflects a lack of novelty rather than superior generative modeling, since SMOTE cannot populate regions of the minority manifold not already spanned by existing samples. Among the deep generative models, the diffusion model produces the most realistic samples by a clear margin, with a C2ST AUC of 0.70 against 0.85 for the cVAE and 0.88 for the cGAN, and a competitive MMD. Figure 3 corroborates this finding: the diffusion-synthesized points (light) closely track the clusters of the real minority samples (dark) in the t-SNE projection [24], covering the same regions while adding plausible intra-cluster variation. Figure 4 summarizes the C2ST comparison.

Table 3. Distributional fidelity of synthetic minority samples

Method	MMD (↓)	C2ST AUC (→0.5)
SMOTE	0.0079 ± 0.0011	0.558 ± 0.022
cGAN	0.0138 ± 0.0039	0.877 ± 0.034
cVAE	0.0112 ± 0.0021	0.849 ± 0.013
Diffusion	0.0127 ± 0.0039	0.700 ± 0.027

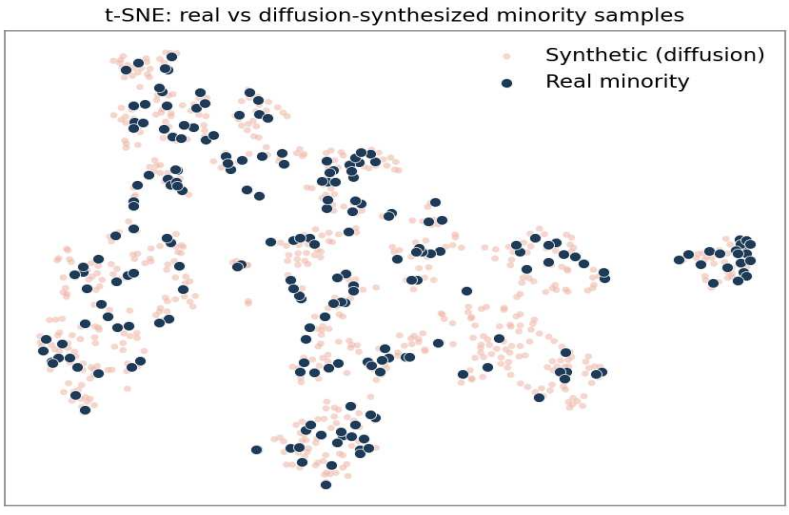


Figure 3. t-SNE projection of the standardized process features for real minority samples and diffusion-synthesized samples

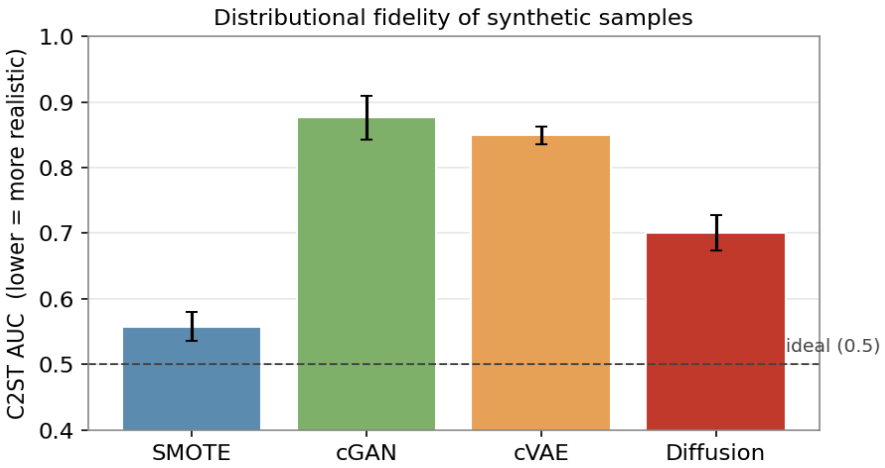


Figure 4. Classifier two-sample test (C2ST) AUC of the synthetic samples.

IV.2 Downstream Diagnostic Performance

Tables 4 and 5 report downstream performance for the MLP and the random forest, respectively. With both classifiers, diffusion augmentation achieves the highest minority-class recall and the highest balanced accuracy. For the MLP, minority recall rises from 0.607 without augmentation to 0.764 with diffusion, and balanced accuracy from 0.684 to 0.803, the gains are even larger for the random forest, whose minority recall climbs from 0.414 to 0.756 and balanced accuracy from 0.531 to 0.799. In both cases the diffusion model surpasses not only the no-augmentation and classical resampling baselines but also the competing generative models. Figure 5 contrasts minority recall across methods and classifiers, and Figure 6 shows balanced accuracy and macro-F1 for the MLP.

Table 4. Downstream performance with the MLP classifier (mean ± std over eight seeds).

Method	Macro F1	Bal. Acc.	Min. Recall	Min. Prec.	Min. F1
No aug.	0.689±0.020	0.684±0.059	0.607±0.074	0.705±0.146	0.613±0.025
Random OS	0.711±0.017	0.727±0.021	0.662±0.027	0.627±0.039	0.641±0.021
SMOTE	0.700±0.022	0.741±0.025	0.680±0.032	0.595±0.051	0.628±0.027
cGAN	0.647±0.041	0.705±0.031	0.636±0.039	0.525±0.094	0.563±0.051
cVAE	0.708±0.017	0.743±0.039	0.682±0.049	0.614±0.026	0.637±0.021
Diffusion	0.636±0.022	0.803±0.036	0.764±0.047	0.447±0.027	0.550±0.027

Table 5. Downstream performance with the random forest classifier (mean ± std over eight seeds).

Method	Macro F1	Bal. Acc.	Min. Recall	Min. Prec.	Min. F1
No aug.	0.604±0.032	0.531±0.039	0.414±0.048	0.683±0.038	0.507±0.040
Random OS	0.647±0.030	0.598±0.034	0.498±0.043	0.682±0.080	0.561±0.037
SMOTE	0.665±0.023	0.672±0.023	0.593±0.029	0.583±0.038	0.584±0.028
cGAN	0.636±0.038	0.698±0.037	0.627±0.046	0.504±0.051	0.548±0.046
cVAE	0.641±0.035	0.664±0.042	0.585±0.052	0.558±0.048	0.555±0.043
Diffusion	0.659±0.019	0.799±0.027	0.756±0.034	0.480±0.024	0.578±0.023

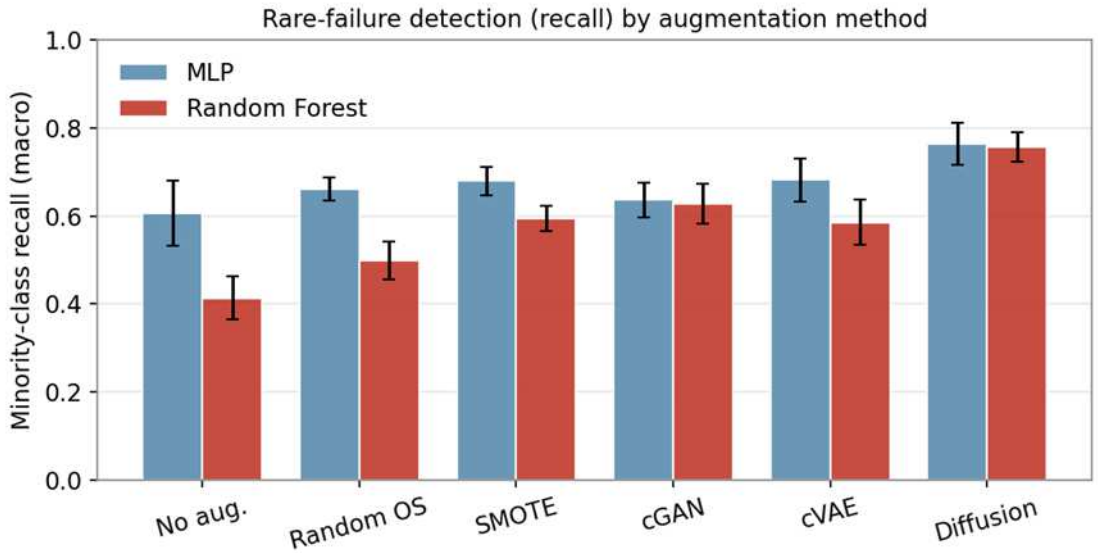


Figure 5. Minority-class recall by augmentation method for both downstream classifiers.

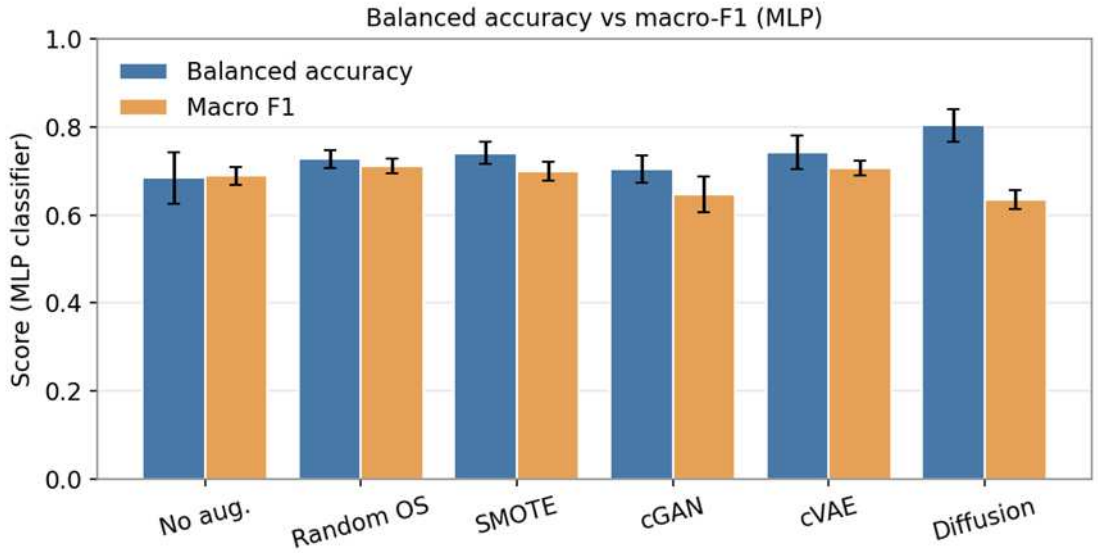


Figure 6. Balanced accuracy and macro-F1 for the MLP classifier.

IV.3 The Precision-Recall Trade-off

The strong recall improvement of diffusion augmentation is accompanied by a reduction in minority-class precision, and consequently in minority and macro-F1. This behavior is coherent: by populating the minority modes with abundant, diverse synthetic samples, diffusion shifts the decision boundary so that more genuine failures are captured, at the cost of additional false alarms on borderline healthy points. SMOTE, random oversampling and the cVAE, which add less diverse samples, retain higher precision but detect fewer real failures. Figure 7 makes the trade-off explicit, placing each method in the recall precision plane. The appropriate operating point is application-dependent: in safety or cost-critical maintenance, where the penalty for a missed failure dominates, the high-recall regime delivered by diffusion is the more

valuable, where false alarms are expensive, a more balanced method such as cVAE may be preferred.

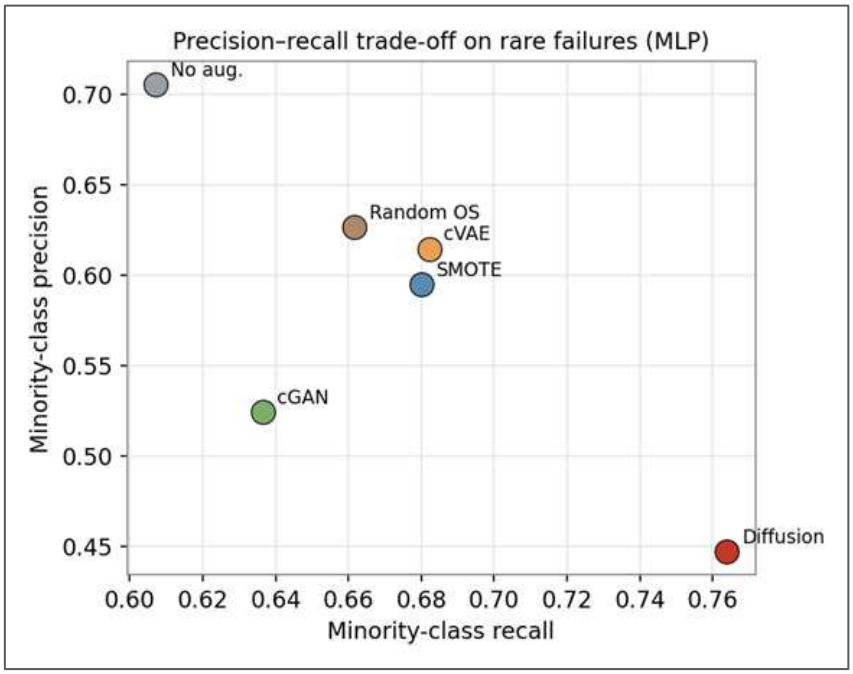


Figure 7. Precision-recall trade-off on the rare failure modes (MLP)

IV.4 Per-Class Analysis

Figures8,9 compares the confusion matrices of the MLP without augmentation and with diffusion augmentation on the reference seed. Without augmentation the rarest mode, TWF, is almost entirely absorbed into the healthy class, and the other modes lose a non-trivial fraction of their true positives to |No Failure|. Diffusion augmentation recovers a substantial share of these missed failures across all four modes, the off-diagonal mass that appears in the healthy row corresponds to the additional false alarms that explain the precision reduction. The qualitative picture is therefore consistent with the aggregate metrics: augmentation moves errors away from the operationally dangerous |missed failure| cell toward the less costly |false alarm| cell.

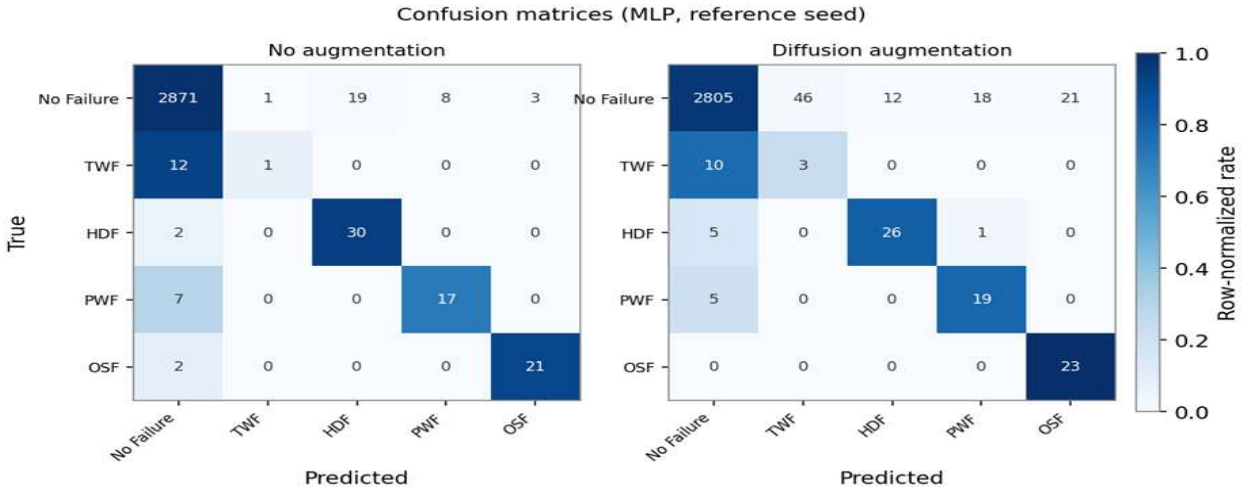


Figure 8. Confusion matrices for the MLP classifier without augmentation

Figure 9. Confusion matrices for the MLP classifier with diffusion augmentation

IV.5 Statistical Significance and Robustness

Across the eight seeds, diffusion augmentation improves minority recall over every baseline in every single run, for both classifiers. The two-sided Wilcoxon signed-rank test confirms that these improvements are significant, returning $p = 0.008$ against each of the five baselines for both the MLP and the random forest the smallest p-value attainable with eight paired observations under a fully consistent direction of effect. Because the conclusion holds for two structurally different classifiers and for all random partitions, the advantage of diffusion augmentation for rare-failure detection is robust rather than incidental to a particular model or split.

IV.6 Per-Class Recovery of Rare Failures

The macro metrics hide useful structure, so Table 6 breaks the minority recall down by failure mode for the MLP. The picture is consistent with the confusion matrices: diffusion augmentation improves recall on every one of the four modes, and it does so most where there is most to gain. Power and overstrain failures, which are tightly tied to the process variables, are recovered almost completely (OSF recall reaches 0.96). Heat-dissipation failure also improves clearly. Tool-wear failure remains the hardest case for all methods its trigger is a randomly drawn wear time only weakly related to the instantaneous features yet even there diffusion roughly quadruples the no-augmentation recall, from 0.06 to 0.21. In other words, augmentation does not paper over an intrinsically unpredictable mode, but it extracts whatever signal is available far better than the baselines.

Table 6. Per-mode minority recall with the MLP classifier (mean over seeds).

Method	TWF	HDF	PWF	OSF
No aug.	0.058	0.820	0.729	0.696
SMOTE	0.115	0.828	0.833	0.891
cVAE	0.212	0.820	0.833	0.880
Diffusion	0.212	0.875	0.865	0.957

IV.7 Sensitivity to the Augmentation Target

How much synthetic data should one add? We swept the per-class target N_target over {250, 500, 1000} while keeping everything else fixed. Figure 10 and Table 7 report the result. Both balanced accuracy and minority recall rise monotonically with the target, from 0.77/0.71 at N_target = 250 to 0.83/0.81 at N_target = 1000. The gain is not free more synthetic minority points push the decision boundary further into healthy territory, so the precision cost discussed in Section (The Precision Recall Trade-off) grows in step but within the range tested, a larger target consistently buys more rare-failure detection. We adopt N_target = 500 in the main experiments as a deliberately moderate operating point, so that the comparison against the baselines is not driven by an unusually aggressive augmentation level.

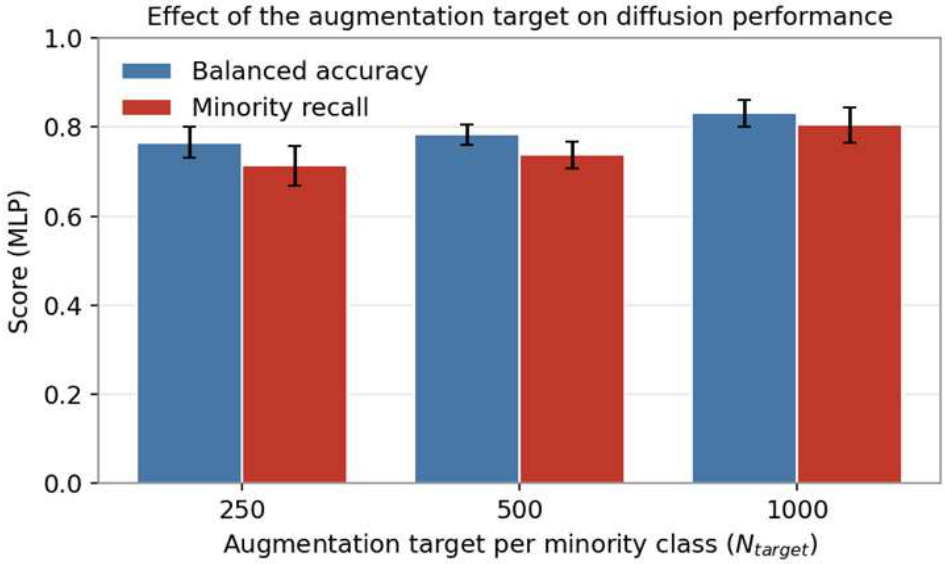


Figure 10. Effect of the augmentation target N_target on diffusion performance (MLP).

Table 7. Diffusion performance as a function of the per-class augmentation

N_target	Balanced accuracy	Minority recall
250	0.765 ± 0.035	0.714 ± 0.045
500	0.783 ± 0.023	0.738 ± 0.030
1000	0.831 ± 0.031	0.806 ± 0.040

IV.8 Sensitivity to the Number of Diffusion Steps

The number of diffusion steps T controls how finely the reverse process is discretized. Sweeping T over {50, 100, 200, 400} exposes a revealing dissociation, shown in Figure 11. Distributional fidelity improves markedly as T grows from 50 to 200 the C2ST AUC falls from 0.81 toward 0.69, i.e. the synthetic samples become harder to distinguish from real ones and then plateaus. Downstream utility, by contrast, is remarkably flat: minority recall stays within a narrow band around 0.73 to 0.77 across the whole range, with no clear monotonic trend. This is the same lesson as the SMOTE comparison, seen from a different angle: beyond a modest number of steps, making the synthetic data still more realistic does little for the classifier, because the useful signal plausible diversity inside each failure mode is already present. A practitioner can therefore use a small T to save sampling time without sacrificing downstream performance; we keep T = 200 in the main study as a conservative, high-fidelity setting.

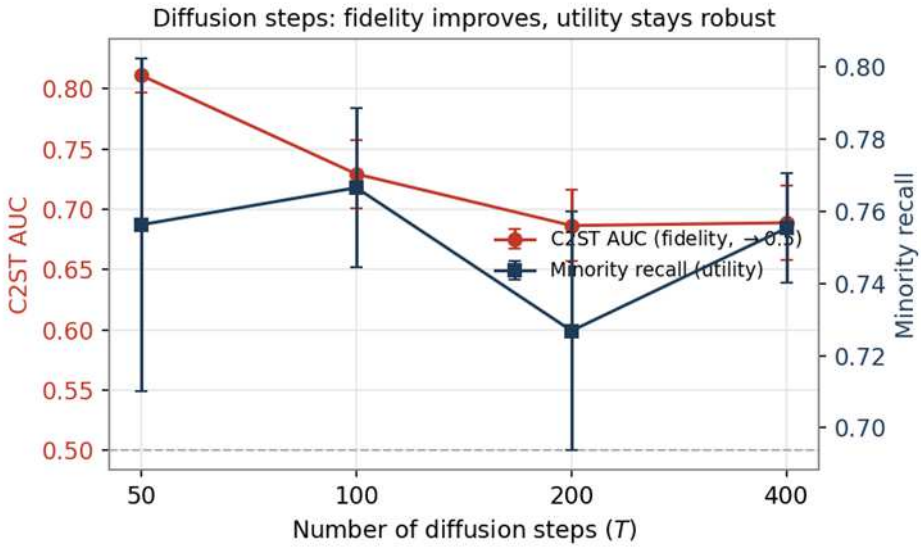


Figure 11. Effect of the number of diffusion steps T

IV.9 Computational Cost

Finally, Table 8 reports the wall-clock time to train each generator and synthesize the augmented set, measured on a single CPU core. The absolute numbers are small tens of seconds at most because both the data and the models are compact, but the ordering is informative. SMOTE is essentially free. Among the deep generators, the diffusion model is cheaper to fit than the cGAN, whose adversarial training is the slowest, and comparable to the cVAE. In other words, the accuracy advantage of diffusion augmentation does not come at a computational premium over the alternatives, which makes it a practical default rather than an expensive luxury.

Table 8. Wall-clock time to train each generator and produce the augmented training set (single CPU core, indicative).

Method	Time (s)
SMOTE	0.1
cVAE	11.7

Diffusion (T=200)	14.9
cGAN	**23.9**

V Discussion

A first lesson from these results is that fidelity and downstream usefulness do not always go together. SMOTE produces the synthetic points that sit closest to the real data, yet it is the diffusion model whose samples are more novel and only moderately distinguishable from real ones that gives the best rare-failure detection downstream. In other words, the most useful augmentation is not the one that copies the training data most faithfully, but the one that adds informative, plausible diversity to the minority manifold. The t-SNE projection in Figure 3 shows diffusion doing exactly that: filling out the failure clusters instead of merely tracing their existing outline.

The metric one reports also has to match the operational objective, and this is easy to get wrong. Looking only at macro-F1 would mislead a maintenance engineer, because F1 penalizes the false alarms that come with high recall and quietly rewards methods that miss more real failures. In condition monitoring the costs are asymmetric a missed failure hurts far more than a false alarm so balanced accuracy and minority recall are the honest indicators, and on both the diffusion model is the clear leader. We still report precision and F1 in full, partly for transparency and partly because a plant with a different cost structure may legitimately weigh them differently.

Finally, the fact that the same conclusion holds for an MLP and for a random forest is reassuring. Two classifiers with very different inductive biases agree, which suggests that diffusion augmentation improves the training distribution itself rather than exploiting a quirk of one model. That matters in practice, where the classifier that ends up deployed is often not the one used during development.

Taken together, the ablations turn these observations into concrete guidance. The augmentation target is the main lever on the recall precision balance: raising it buys more rare-failure detection but spends precision, so it should be set from the cost of a missed failure versus a false alarm in the specific plant, not chosen blindly. The number of diffusion steps, by contrast, is not a quality knob for the downstream task a small value is enough so it can be set from the available sampling budget. And because diffusion is no more expensive to train than the generative baselines, adopting it carries little practical overhead. A reasonable default recipe for a new asset, then, is to start from a moderate target, monitor minority recall and precision on a held-out real set, and increase the target only as far as the tolerated false-alarm rate allows.

VI Limitations and Threats to Validity

Several caveats temper these conclusions. The AI4I 2020 dataset is synthetic, and although it is engineered to reflect realistic industrial behavior, validation on field data from operating machinery remains necessary before operational deployment real condition-monitoring data bring sensor noise, drift and labeling uncertainty that a simulator does not fully capture. The data are also tabular and snapshot-based, whereas much industrial monitoring produces multivariate time series. Extending the diffusion augmenter to sequences, where temporal structure and inter-sensor correlations must be preserved, is an important and non-trivial next step, since a naive per-timestamp model would destroy exactly the dynamics that distinguish a developing fault.

The precision reduction that accompanies aggressive augmentation is acceptable in safety critical contexts but must be controlled where false alarms are themselves costly the augmentation target and a tuned decision threshold are the natural levers, and our ablation

shows the target behaves predictably enough to be set deliberately. We also fixed the generator and classifier hyper-parameters a priori rather than tuning them per method this protects the comparison from over-fitting to the test set but means the absolute numbers are conservative, and a careful per-method search would likely raise every generative method somewhat. Our significance testing rests on eight paired seeds, which is sufficient to establish a consistent direction of effect but is a small sample the perfectly consistent ranking across seeds and across two classifiers nonetheless makes a spurious conclusion unlikely. Finally, we evaluated a single dataset and two classifier families: the qualitative story that diffusion maximizes rare-failure recall at a transparent precision cost should be re-examined on other benchmarks and learners before being treated as general.

VII Conclusion and Future Work

This paper presented a reproducible benchmark of conditional generative data augmentation for imbalanced multi-mode fault diagnosis in predictive maintenance, centered on a class-conditional denoising diffusion model and evaluated on the AI4I 2020 dataset. Under an identical protocol, diffusion augmentation delivered the highest rare-failure recall and balanced accuracy with two independent classifiers, improving minority recall by up to 0.34 in absolute terms over the no-augmentation baseline and outperforming random oversampling, SMOTE, a conditional GAN and a conditional VAE, with the advantage significant and consistent across all eight runs. A precision recall trade-off was characterized openly, situating diffusion augmentation as the method of choice when the operational priority is to minimize missed failures. Future work will extend the approach to time-series condition-monitoring signals [20] and real field data, integrate cost-aware calibration to manage the false-alarm rate, and study guided and classifier-free diffusion variants to control the fidelity diversity balance of the synthetic failure data.

REFERENCES

1. Zhang, W., Yang, D., Wang, H.: Data-Driven Methods for Predictive Maintenance of Industrial Equipment: A Survey. IEEE Systems Journal. 13, 2213–2227 (2019). https://doi.org/10.1109/JSYST.2019.2905565.

2. Çınar, Z.M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M., Safaei, B.: Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Sustainability. 12, 8211 (2020).

3. Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R. da P., Basto, J.P., Alcalá, S.G.: A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering. 137, 106024 (2019).

4. Hamdi Alaoui, A., Meddaoui, A., Hain, M.: AI-driven predictive maintenance for industry 4.0: a systematic review of models, methods, and challenges. Int J Adv Manuf Technol. 143, 4861–4876 (2026). https://doi.org/10.1007/s00170-026-17531-w.

5. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: TabDDPM: Modelling Tabular Data with Diffusion Models, http://arxiv.org/abs/2209.15421, (2024). https://doi.org/10.5555/3618408.3619133.

6. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 5, 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0.

7. Kingma, D.P., Welling, M.: An Introduction to Variational Autoencoders. Foundations and Trends in Machine Learning. 12, 307–392 (2019). https://doi.org/10.1561/2200000056.

8. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics, http://arxiv.org/abs/1503.03585, (2015). https://doi.org/10.48550/arXiv.1503.03585.

9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16, 321–357 (2002). https://doi.org/10.1613/jair.953.

10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks, http://arxiv.org/abs/1406.2661, (2014). https://doi.org/10.48550/arXiv.1406.2661.

11. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models, http://arxiv.org/abs/2006.11239, (2020). https://doi.org/10.48550/arXiv.2006.11239.

12. Lei, Y., Li, N., Guo, L., Li, N., Yan, T., Lin, J.: Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical Systems and Signal Processing. 104, 799–834 (2018). https://doi.org/10.1016/j.ymssp.2017.11.016.

13. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular data using Conditional GAN, http://arxiv.org/abs/1907.00503, (2019). https://doi.org/10.48550/arXiv.1907.00503.

14. Matzka, S.: Explainable Artificial Intelligence for Predictive Maintenance Applications. 2020 Third International Conference on Artificial Intelligence for Industries (AI4I). 69–74 (2020). https://doi.org/10.1109/AI4I49448.2020.00023.

15. Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., Xu, E.: Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. ISA Transactions. 119, 152–171 (2022). https://doi.org/10.1016/j.isatra.2021.02.042.

16. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A Kernel Two-Sample Test. Journal of Machine Learning Research. 13, 723–773 (2012).

17. He, H., Garcia, E.A.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 21, 1263–1284 (2009). https://doi.org/10.1109/TKDE.2008.239.

18. Review of imbalanced fault diagnosis technology based on generative adversarial networks | Journal of Computational Design and Engineering | Oxford Academic, https://academic.oup.com/jcde/article/11/5/99/7749581, last accessed 2026/06/23.

19. Sun, S., Kolar, M., Xu, J.: Learning structured densities via infinite dimensional exponential families. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2015).

20. Yi, H., Hou, L., Jin, Y., Saeed, N.A., Kandil, A., Duan, H.: Time series diffusion method: A denoising diffusion probabilistic model for vibration signal generation. Mechanical Systems and Signal Processing. 216, 111481 (2024). https://doi.org/10.1016/j.ymssp.2024.111481.

21. Zhang, X., Qin, Y., Yuen, C., Jayasinghe, L., Liu, X.: Time-Series Regeneration With Convolutional Recurrent Generative Adversarial Network for Remaining Useful Life Estimation. IEEE Transactions on Industrial Informatics. 17, 6820–6831 (2021). https://doi.org/10.1109/TII.2020.3046036.

22. Unknown: AI4I 2020 Predictive Maintenance Dataset, https://archive.ics.uci.edu/dataset/601, (2020). https://doi.org/10.24432/C5HS5C.

23. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1, 80–83 (1945). https://doi.org/10.2307/3001968.

24. Maaten, L. van der, Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research. 9, 2579–2605 (2008).

Statements & Declarations

Funding :

The authors declare that no funds, no data set used, grants, or other support were received during the preparation of this manuscript.

Competing Interest:

Authors of this paper have no financial interests and have no relevant financial or non-financial interests to disclose."

Author Contributions:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Hamdi Alaoui. A, Meddaoui.

The first draft of the manuscript was written by Mr. Hamdi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.