Mars: Jurnal Teknik Mesin, Industri, Elektro dan Ilmu Komputer
Volume 3, Nomor 6, Desember 2025
e-ISSN : 3031-8742, p-ISSN : 3031-8750, Hal. 168-183
DOI: https://doi.org/10.61132/mars.v3i6.1262
Tersedia: https://journal.arteii.or.id/index.php/Mars

Bridging The Synthetic-To-Real Gap: A Model-Data Coevolution Approach With Stochastic Feature Decoupling For Ac Unit Fault Diagnosis

Muhamad Raynard Alif[1], Mukhammad Andri Setiawan[2]
[1-2] Prodi Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia
**Penulis Korespondensi: 20523167@students.uii.ac.id*

Abstract. *The scarcity of real-world data is a major challenge in the development of machine learning-based air conditioning (AC) fault diagnosis systems, making the use of synthetic data an unavoidable necessity. However, synthetic datasets generated through rule-based approaches often suffer from a significant simulation-to-real gap, negatively impacting the model's generalizability. To address this issue, we propose a Model-Data Coevolution (MDC) framework that simultaneously optimizes the model and data generation processes. MDC leverages a Simulated Annealing (SA) controller to adaptively optimize data augmentation parameters based on model performance during training. Furthermore, we introduce a novel augmentation technique called Stochastic Feature Decoupling (SFD). Unlike traditional Logical Consistent Augmentation (LCA) approaches that maintain deterministic relationships between features, SFD applies stochastic noise independently to both raw and derived features. This approach intentionally breaks the rigid physical relationships between features to encourage the model to learn a more robust representation. The empirical evaluation results show that SFD significantly outperforms LCA, achieving a weighted F1 score of 0.93 and increasing recall in the NORMAL class by up to 82%. These findings demonstrate that, despite generating physically "impossible" samples, SFD serves as a robust form of regularization. Thus, this approach is capable of improving model generalization in the face of real-world complexity and uncertainty, particularly in synthetic data-driven AC fault diagnosis systems.*

Keywords: *Model-Data Coevolution, Synthetic Data, Domain Gap, Sim-to-Real, Fault Detection and Diagnosis (FDD).*

1. INTRODUCTION

The Need for Reliable Fault Diagnosis

Heating, Ventilation, and Air Conditioning (HVAC) systems are a critical and energy-intensive component of the modern built environment. As a popular research topic, their optimization is central to global energy management, as buildings' energy is predominantly allocated to heating and cooling [1]. In commercial buildings, HVAC systems can consume an average of 52% of total energy used [2], with some analyses placing their responsibility between 40% and 80% of a building's energy consumption [3]. This immense energy footprint has significant environmental consequences, the air conditioning industry alone is estimated to be responsible for approximately 4% of global greenhouse gas emissions annually [2], positioning HVAC management as a key factor in decarbonization efforts [4].

The operational state of these systems is a critical lever for energy efficiency. The prevalence of undetected operational faults such as failed damper actuators, sensor drift, or improper control sequences [5], directly exacerbates energy consumption and system unreliability. Various HVAC faults can lead to energy losses of up to 20% of a building's total energy consumption [6]. More broadly, system faults and inefficient controls are estimated to

Naskah Masuk: 04 November 2025; Revisi: 30 November 2025; Diterima: 27 Desember 2025;
Terbit: 30 Desember 2025

cause energy wastage of 15% to 30% in buildings [4]. These faults are pervasive, a 2023 empirical study analyzing over 60,000 pieces of HVAC equipment found that on any given day, 40% of air handling units (AHUs) and 30% of air terminal units experienced a reported fault of some kind [5].

In response, the field has increasingly shifted from traditional, reactive diagnostics to to automated (AFDD) and data-driven solutions [8]. The integration of advanced analytics, machine learning (ML), and artificial intelligence (AI) has gained considerable attention [10]. These methods offer a pathway to enhance diagnostic accuracy, reduce reliance on deep expert knowledge often required by physicsbased models [1], and enable the proactive, continuous monitoring necessary to minimize the time a building spends in a faulty operational state [6]. The ultimate goal is to enhance system reliability and energy conservation, with data-driven FDD being the most promising vector for achieving this at scale.

The Data Scarcity Challenge

The consensus on data-driven FDD as the preferred path forward is predicated on the power of machine learning (ML) and deep learning (DL) models [2]. These algorithms have demonstrated high accuracy in capturing the complex, non-linear relationships inherent in HVAC operational data [11]. However, the efficacy of these models, particularly in supervised or semi-supervised contexts, is critically dependent on the availability of large, diverse, and welllabeled training datasets [12]. For a model to achieve effective and accurate diagnosis in a real-world scenario, it must be trained on a substantial number of categories with a balanced distribution of sample sizes.

This requirement creates a central paradox. Modern facilities, equipped with Building Automation Systems (BAS) and Building Energy Management Systems (BEMS), are in an era of "big data," generating enormous volumes of sensor measurements [1]. Yet, the FDD field is simultaneously plagued by a critical "data scarcity" [10]. This scarcity does not refer to the volume of data, but to the availability of labeled fault data. This deficit is a direct consequence of engineering reliability. HVAC systems are designed to operate in a fault-free state, consequently, faulty data samples are inherently more difficult to collect than normal operation data [13]. The result is a dataset landscape defined by what it lacks. This imbalance, where normal data vastly overwhelms the rare fault instances, is a "major challenge" that can lead to "highly biased classifiers" that are effectively blind to the minority fault classes [15].

The seemingly straightforward solution to manually collect and label more real-world fault data is operationally and economically non-scalable. Such datasets are "extremely difficult to come by" [16]. The process of inducing faults for data collection can be disruptive

and costly, and the subsequent data annotation process is a significant bottleneck that requires deep domain expertise [16]. The cost associated with the required sensors and expert-hours can be prohibitive, particularly in the residential sector [7]. This "scarcity of labelled datasets" and the general lack of reliable, audited public data constitute the primary roadblock to the widespread, effective deployment of ML-based FDD [15].

Sim-to-Real Domain Gap

Given the intractability of acquiring sufficient real-world fault data, the research community has logically pivoted to a seemingly ideal alternative, synthetic data generation [17]. Using physics-based simulation, researchers can generate vast, abundant, and, crucially, perfectly labeled datasets encompassing a wide array of fault types and severities [18]. This approach solves the scarcity and imbalance problems at a stroke, enabling the training of complex, data-hungry models [19]. "However, this strategy introduces the formidable 'sim-to-real' domain gap [20], a barrier arising from discrepancies between simulated and physical environments [21]. Models trained exclusively on idealistic synthetic data tend to overfit to deterministic simulation rules [23], consequently failing when deployed on real-world data that diverges from the training distribution [22].

The fundamental cause of this domain gap lies in the difference between the data-generating processes of simulated and real-world environments. While physics-based simulations are comprehensive, they inherently fail to capture the full range of stochastic, unmodeled variability present in real-world operations [24]. Synthetic data oversimplifies reality by lacking complex artifacts such as sensor noise, machine wear, environmental fluctuations, thermal noise, sensor drift, and electrical interference [25].

This failure to generalize is not merely a lack of robustness; it is a systematic failure mode of machine learning. The "clean" synthetic data is rife with "spurious correlations", non-predictive features that are artifacts of the simulation [26]. The learning models, which are susceptible to "simplicity bias," will preferentially learn these simple, spurious features rather than the complex, invariant features that are truly predictive of a fault. When the model is deployed, these spurious correlations no longer hold, leading to diagnostic failure. The field is thus at an impasse, having traded the problem of data *scarcity* for the problem of data *fidelity*. We demonstrate the severity of this gap with our own baseline experiment. A Random Forest model was trained on the 'clean' expert-defined synthetic dataset and then validated against our real-world dataset. The results, shown in Figure 1, are a practical failure. While the model performs perfectly on synthetic data, it misclassifies over 25% (1,347 samples) of real-world NORMAL operations as TROUBLE. This 'crying wolf'

phenomenon renders the model unusable in a production environment, as it would flood maintenance logs with false positives and destroy operator trust.

Figure 1. Baseline model (trained on clean synthetic data) failure on real-world validation set.

Model-Data Coevolution (MDC)

The persistence of this sim-to-real gap, as demonstrated in Figure 1, suggests a fundamental limitation in the prevailing "Model-Centric" research paradigm. This paradigm, which treats data as a static asset and focuses on refining algorithms, has failed to solve the data-fidelity problem. The Data-Centric paradigm posits that "systematically designing datasets" and "engineering data quality" are the most critical and effective levers for improving the performance, robustness, and fairness of AIbased systems.

This research *implements* this philosophy through a Model-Data Coevolution (MDC) system. The concept of data and models evolving in tandem has recently emerged as a powerful paradigm for solving complex, real-world problems. In our MDC system, the synthetic dataset is not a static artifact but an "organism" that is iteratively improved. The most obvious approach is LogicallyConsistent Augmentation (LCA), where noise is added to raw sensor values and the scaled features are perfectly recalculated.

However, we hypothesize this is *still* too "clean." We propose a counter-intuitive alternative, the Stochastic Feature Decoupling (SFD). A novel technique that applies independent, stochastic noise to both the raw and scaled features simultaneously. This intentionally "decouples" their deterministic link, forcing the model to learn a more robust, non-linear relationship that, we argue, better reflects realworld imperfections. This paper presents a complete MDC framework and provides empirical A/B comparison between the LCA and SFD augmentation strategies. We will demonstrate that, counter-intuitively, the SFD

method produces a model with superior generalization and resolves the critical "crying wolf" misclassification.

2. METHODOLOGY

Datasets and Preprocessing

Our methodology relies on two key datasets, a baseline synthetic dataset for training and a real-world dataset for validation.

1. Baseline Synthetic Dataset

The baseline synthetic data was generated using a custom, stateful, rule-based simulator. This simulator models 12 distinct AC unit types under 7 operational conditions (Normal, Maintenance 1, Maintenance 2, Abnormal, Trouble 1, Trouble 2, Trouble 3). This initial dataset contains 42,000 samples and represents the "clean," idealized data that we aim to evolve.

2. Real-World Validation Dataset

The validation dataset consists of 9,169 sensor measurements collected from real-world, in-field AC units. This "ground truth" was established by applying the same expert-defined rule-based logic to the raw sensor readings. This distinction is critical, the "sim-to-real" gap we are addressing is *not* a difference in labeling logic, but purely a difference in the feature distributions between the "clean" synthetic sensor readings and the "noisy" real-world sensor readings. This dataset is limited to the three conditions observed in the field, NORMAL (5,598 samples), Maintenance 1 (2,412 samples), and Maintenance 2 (1,159 samples). This dataset is held out and used exclusively by the validator component to score the fitness of each evolved model.

3. Feature Engineering and Scaling (Tiered-Scaling)

A critical aspect of our data is the feature set. Both synthetic and real dataset were created using the same preprocessing script, which centralizes its logic in scaler. This script generates two "views" of the data, which are both fed to the model:

- Raw Features. The direct sensor values.
- Scaled Features. A set of expert-defined, tieredscaled features, where NORMAL values are mapped to [0, 1], MAINTENANCE values mapped to [-1, 0] and [1, 2], and TROUBLE values mapped to [<-1] and [>2]. Our model is trained on *both* sets of numeric features simultaneously. The core of our experiment tests the *relationship* between these two feature streams during augmentation.

The Model-Data Coevolution (MDC) Framework

Our system is a closed-loop framework designed to iteratively evolve the baseline synthetic dataset to produce a model that performs better on the real validation dataset.

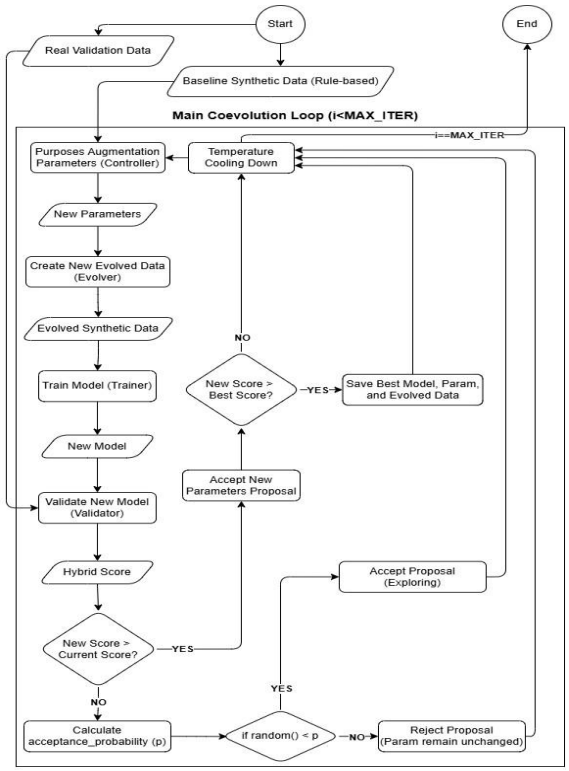

Figure 2. Flowchart of Model-Data Coevolution loop

Our system is a closed-loop framework designed to iteratively evolve the baseline synthetic dataset to produce a model that performs better on the real validation dataset.

1. The Controller: The "brain" that proposes new augmentation parameters.
2. The Evolver: The "hands" that take the parameters and create a new, "evolved" synthetic dataset from the synthetic data baseline.
3. The Trainer: The "student" that trains a new Random Forest model on the evolved dataset.
4. The Validator: The "judge" that scores the new model against real validation data and provides a "fitness score" back to the Controller.

Component Implementation

1. The Trainer is a Random Forest Classifier from the scikit-learn library, we use n_estimators=250, min_samples_split=4, and criterion='entropy'. Critically, class_weight='balanced' is used to manage the inherent class imbalance in the 7 condition synthetic training data. The model is trained on all available numeric features (both raw and scaled) after NaN values are imputed with 0.

2. The Validator & Objective Function. The Validator's role is to provide a single fitness score for the Controller. It does this by testing the trained model against the held-out real validation data. The fitness is determined by a hybrid score, which is a weighted average of model performance and data divergence, defined as:

$Composite = (0.2 \times F1) + (0.3 \times P) + (0.5 \times R)$

$Hybrid = (0.9 \times Composite) + (0.1 \times (1 - Div))$

Where *F1*, *P*, and *R* are the weighted average F1score, Precision, and Recall from the validation report. Div is the feature divergence score, a simple metric (mean normalized difference between the real and synthetic feature means) used to penalize evolved datasets that drift too far from the real data's distribution.

3. The Controller or the "brain" of the operation is a Simulated Annealing (SA) algorithm.

SA is a metaheuristic optimization algorithm chosen for its ability to escape local optima. A simple hillclimbing algorithm would get "stuck" on the first good-but-not-great solution. SA avoids this by always accepting a *better* score, but also *sometimes* accepting a *worse* score with a probability (*p*) defined by the current "temperature" (*T*):

$$p = e\left(\frac{scorenew - scorecurrent}{T}\right)$$

As the iterations proceed, the temperature T is slowly lowered via a cooling rate, making the algorithm gradually "settle" on the best-found solution.

4. Augmentation Parameter Space. The Controller's task is to find the optimal set of augmentation parameters. These parameters are fed to the evolver to evolve the baseline dataset. This parameter space consists of four key hyperparameters:

a. Rule Exponent (rule_exp): A multiplier applied to all raw and scaled sensor values before noise is added. It "stretches" or "shrinks" the entire dataset's value range. A value > 1.0 exaggerates features, while a value < 1.0 mutes them.

- Range: [0.5, 2.0]

b. Fault Intensity / Noise (fault_int): The primary amplitude for the stochastic noise (np.random.normal). A higher value creates *more* noise and "messier" data, while a lower value results in a "cleaner" dataset closer to the baseline.

- Range: [0.5, 2.0]

c. Relative Frequency (rel_freq): A parameter intended to control the relative frequency of fault conditions (this inside the controller, but its effect in evolver is is simplified to a multiplier, similar to rule_exp).

- Range: [0.5, 2.0]

d. Thermal Drift (thermal): A small, additive value (±0.05) intended to simulate minor sensor drift or thermal-related offsets.

- Range: [-0.5, 0.5]

Experimental Design: Augmentation Strategies

1. Experiment A: Stochastic Feature Decoupling (SFD) This method, implemented in our original synthetic data evolver, tests the hypothesis that *decoupling* the feature streams is a more realistic augmentation. It applies independent, stochastic noise ($\in_1$, $\in_2$) to both the raw feature stream (X_{raw}) and the scaled feature stream (X_{Scaled}) simultaneously:

$$Xraw' = Xraw \times (param) + \in 1$$

$$Xscaled' = Xscaled \times (param) + \in 2$$

This intentionally breaks the deterministic link

($X_{scaled} \neq f(X_{raw})$), forcing the model to learn from "contradictory" or "messy" data, which we posit is a closer simulation of real-world noise.

2. Experiment B: Logically-Consistent Augmentation (LCA)

This method serves as our control, representing the "common sense" approach. It applies stochastic noise ($\in_1$) only to the raw feature stream (X_{raw}). The original X_{scaled} columns are dropped, and the tiered scaler function (the same function used to create the data) is re-applied to the evolved raw data to generate a new, perfectly consistent $X_{scaled}{}'$:

$$Xraw' = Xraw \times (param) + \in 1$$

$$Xscaled' = f(Xraw')$$

This method ensures the data's logical integrity is *always* preserved.

Coevolution Run Configuration

Both experiments (A and B) were initiated from iteration 1 using the controller settings (start_temperature=0.2, cooling_rate=0.98, random_state=42). The loop was set to run for MAX_ITER = 210 or until EARLY_STOP_PATIENCE = 70 (70 consecutive iterations with no improvement to the all-time best score) was triggered. All code was executed in a Python 3 environment.

3. RESULT

Coevolution Process Comparison

The Model-Data Coevolution (MDC) loop was executed for both experiments. The controller was run for both experiments. The resulting evolution of the best hybrid score is plotted in Figure 3. The two methods demonstrate a stark divergence in optimization potential.

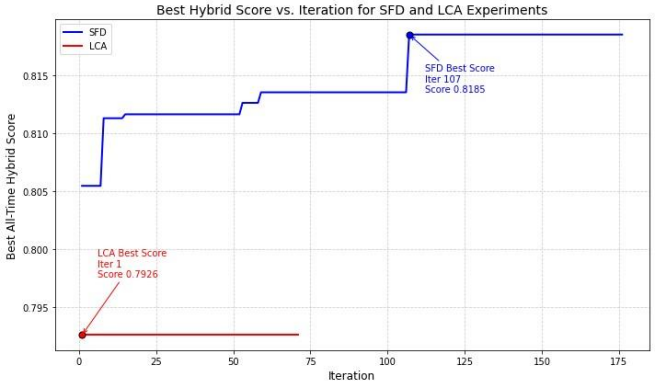

Figure 3. Best Hybrid Score vs. Iteration for SFD and LCA Experiments

1. Experiment A (SFD): The controller ran for 176 iterations before the EARLY_STOP_PATIENCE of 70 iterations was met. The best solution, with a hybrid score of 0.818, was found at iteration 107.
2. Experiment B (LCA): The controller ran for 71 iterations. The best solution, with a hybrid score of 0.792, was found at iteration 4.

The LCA (control) method failed to optimize. The controller found its best hybrid score at Iteration 1 and did not find a superior set of parameters for the subsequent 70 iterations, triggering an early stop. In sharp contrast, the SFD method showed a successful coevolution run. The optimization "climbed" through multiple stages, finding new "best" scores at iterations 8, 15, 53, 59, and finally at Iteration 107, where it converged on a final, superior hybrid score. The final optimized parameters reflect this fundamental difference in strategy:

1. Exp. A (SFD) Best Parameters (Iter 107):
   - rule_exp: 1.00 (No change)
   - fault_int: **0.55** (Low noise)
   - rel_freq: 0.57 (Low frequency)
   - thermal: 0.06 (Small positive drift)
2. Exp. B (LCA) Best Parameters (Iter 1):
   - rule_exp: 0.97 (Slight shrink)
   - fault_int**: 1.09** (High noise)

- rel_freq: 1.05 (High frequency)
- thermal: 0.00 (No change)

The "common sense" LCA experiment immediately failed with a high-noise strategy. The proposed SFD experiment succeeded by learning to apply a *moderate, low-noise* level of "contradiction" to the data.

Final Model Validation

The primary evidence of success is the performance of the final, best-evolved model from each experiment. The classification reports (Figure 4 and 5) and their corresponding confusion matrices (Figure 6 and 7) are presented for direct comparison.

[Validator] Real F1_weighted=0.903

	precision	recall	f1-score	support
ABNORMAL	0.00	0.00	0.00	0
MAINTENANCE 1	0.92	0.99	0.96	2412
MAINTENANCE 2	0.85	1.00	0.92	1159
NORMAL	1.00	0.78	0.88	5598
TROUBLE 1	0.00	0.00	0.00	0
TROUBLE 2	0.00	0.00	0.00	0
TROUBLE 3	0.00	0.00	0.00	0
accuracy			0.86	9169
macro avg	0.40	0.40	0.39	9169
weighted avg	0.96	0.86	0.90	9169

Figure 4. Classification Report for the Best Model from Experiment A (SFD

[Validator] Real F1_weighted=0.931

	precision	recall	f1-score	support
ABNORMAL	0.00	0.00	0.00	0
MAINTENANCE 1	0.96	1.00	0.98	2412
MAINTENANCE 2	0.94	1.00	0.97	1159
NORMAL	1.00	0.82	0.90	5598
TROUBLE 1	0.00	0.00	0.00	0
TROUBLE 3	0.00	0.00	0.00	0
accuracy			0.89	9169
macro avg	0.48	0.47	0.47	9169
weighted avg	0.98	0.89	0.93	9169

Figure 5. Classification Report for the Best Model from Experiment B (LCA).

The reports show the SFD model is superior in weighted avg F1-score (0.93 vs. 0.90) and NORMAL class recall (0.82 vs. 0.78). The confusion matrices visually confirm this. Figure 6 (SFD) shows 4594 correct NORMAL classifications and a reduction of NORMAL-to-TROUBLE leaks to only 570 samples. Critically, it successfully mitigated the "crying wolf" problem identified in Figure 1. The NORMAL class recall improved to 82.0%.

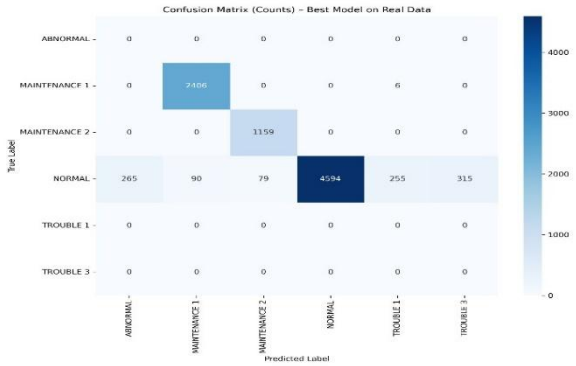

Figure 6. Confusion Matrix (Counts) for the Best Model from Experiment A (SFD).

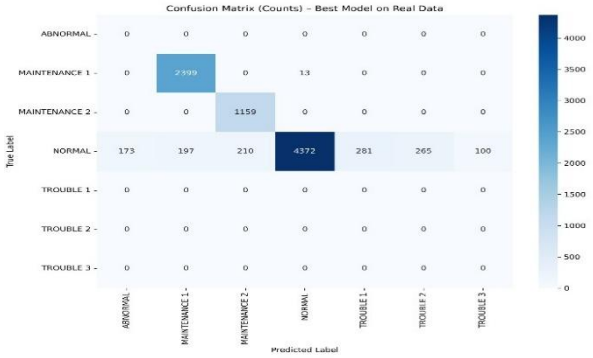

Figure 7. Confusion Matrix (Counts) for the Best Model from Experiment B (LCA).

Figure 7 (LCA) shows that only 4372 correct NORMAL classifications and a larger leak of 646 samples to the TROUBLE classes. While an improvement over the baseline, its performance was demonstrably inferior to the SFD model. The NORMAL class recall was only 78.1%.

Summary of Findings

The results of the entire investigation are summarized in Table 1. The data provides a clear and empirical validation of our hypothesis. The model trained using the proposed Stochastic Feature Decoupling (SFD) method is superior to both the baseline model and the control (LCA) model across all key performance and problem-specific metrics.

Table 1. Comparative summary of model performance against the real-world validation set

Performance Metric	Baseline Mode	LCA Model	SFD Model
Weighted F1Score	0.84	0.90	0.93
Normal Class Recall	0.74	0.78	0.82
Normal Class F1-Score	0.85	0.88	0.90
Normal Leaks To Trouble	1,347 samples	646 samples	570 samples
Best Hybrid Score	N/A	0.7926	0.8185

**Note: Baseline scores are calculated from Figure 1. LCA and SFD scores are from their respective best-performing iterations.*

4. DISCUSSION

Why Did "Impossible" Data Create a Better Model?

The central finding of this paper is that the "logicallyflawed" SFD method produced a"factually-superior" model. This result, while counter-intuitive, provides a deep insight into the nature of the sim-to-real gap. The baseline model's failure (Figure 1) proved that the problem was a *feature distribution gap*. The "clean" synthetic features were simply too different from the "noisy" real-world features. Our experiment set out to close this gap by adding noise.

1. Experiment B (LCA) failed because it was *too logical*. It added noise to the raw features (X_{raw}) but then perfectly "fixed" the scaled features ($X_{scaled}{}' = f(X_{raw}{}')$). This created a new, *noisier*, but still *perfectly logical* dataset. This "book-smart" data was still too idealized and failed to prepare the model for the real world, as shown by its flatlining optimization curve (Figure 3) and inferior NORMAL class recall (78%).
2. Experiment A (SFD) succeeded *because* it was "illogical." By applying independent noise to both X_{raw} and X_{scaled}, it broke the deterministic link between them. It created "impossible" data where Current_scaled was no longer the perfect, mathematical function of Current (A). This "impossible" data was, in fact, the key. It acted as a powerful form of implicit regularization.

The real world is not logically consistent, sensors drift, noise is non-uniform, and the expert-defined rules are an *idealization*, not a law. The real X_{raw} and X_{scaled} are *also* partially decoupled by this real-world noise. The SFD method accidentally created a dataset that was a *far more realistic simulation of the real world's "messiness"* than the "correct" LCA method did. The SFD-trained model became "street-smart." It was forced to learn a robust, flexible relationship between the two feature streams, making it less brittle and far more effective at generalizing to the noisy, imperfect real data.

Analysis of Coevolution Dynamics (The "Smoking Gun")

The optimization plot (Figure 3) and the final parameters (Section III.A) provide the "smoking gun" that proves this interpretation. The LCA controller (Exp. B) was given only one tool, "noise level." It tried a high-noise strategy (fault_int: 1.09), which failed, and it never improved. Its optimization landscape was "flat" or "convex", there was no path to a better

solution. The SFD controller (Exp. A), however, had a second, more powerful tool, the "contradiction level." The plot shows it successfully climbing, finding new peaks at iterations 8, 15, 53, and 107. This proves it was navigating a rich and complex optimization landscape. Most importantly, it converged on a low-noise regime (fault_int: 0.55). This tells a clear story, the controller learned that *a little bit* of contradiction was highly beneficial, but *too much* (like the high-noise strategy in Exp. B) was bad. It successfully "tuned" the level of logical contradiction to find a "sweet spot" of realism, a task the "logically-correct" LCA controller was fundamentally incapable of performing.

Connection to Broader Machine Learning Concepts

This finding, while discovered accidentally, is not a fluke. It aligns strongly with established concepts in machine learning.

1. Data-Centric AI: This research is a clear implementation of the "Data-Centric AI" paradigm. We achieved 8% relative improvement in NORMAL class recall (from 74% to 82%) not by changing the model, but by systematically engineering the quality and nature of the training data.
2. Structured Regularization: The success of SFD can be framed as a novel form of structured regularization. Dropout, for example, is a popular technique that prevents neural networks from "coadapting" by randomly dropping neurons. Our SFD method acts as a feature-level dropout, preventing the model from "co-adapting" to the spurious, perfect correlation between X_{raw} and X_{scaled}.
3. Sim-to-Real Transfer: This work is a direct contribution to the sim-to-real field. While many approaches focus on complex domain adaptation or randomizing simulation physics [20], our work proposes a simpler, data-space augmentation that simulates the imperfection of the real world by breaking the idealized logic of the simulation.

Limitations and Future Work

While the results are strong, this study has limitations that open clear avenues for future research.

1. Model Generalizability: The experiment was conducted using only a Random Forest classifier. It is unknown how more complex models, such as deep neural networks, would react to this "contradictory" SFD data. They might be *more* robust to it, or they might fail to converge.
2. Dataset Generalizability: The SFD method was tested on a single FDD dataset. Its efficacy should be validated on other sim-to-real problems, especially in other industrial domains where sensor data and derived features are common.

3. Parameter Space: The augmentation parameter space was simple. Future work could explore more complex decoupling strategies, such as applying different noise distributions or amplitudes to each feature stream independently.

5. CONLUSION

This paper addressed the critical sim-to-real domain gap in HVAC fault diagnosis, where baseline models trained on "clean" synthetic data fail to generalize to "noisy" real-world data, resulting in a high rate of "crying wolf" errors. We proposed and built a complete Model-Data Coevolution (MDC) system to solve this problem by iteratively evolving the synthetic training data. The core of our research was a comparative experiment between two augmentation strategies, a "common sense" Logically-Consistent Augmentation (LCA) and our proposed, counter-intuitive Stochastic Feature Decoupling (SFD) method, which intentionally breaks the mathematical link between raw and scaled features.

The results were conclusive. The LCA (control) method failed to optimize, finding no improvement beyond its initial parameters. The SFD (proposed) method, however, successfully optimized over 107 iterations, converging on a model that was demonstrably superior. Our final SFDevolved model achieved a weighted F1-score of 0.93 (compared to 0.90 for LCA) and, most importantly, solved the primary "crying wolf" problem by improving NORMAL class recall from 74% to 82.0% (see Table 1). We conclude that the SFD method's success is due to its function as a powerful implicit regularizer. By training on "messy" and "contradictory" data, the model was forced to learn a more robust representation of the features, making it "street-smart" for the imperfections of the real world. This work empirically demonstrates that for bridging the sim-toreal gap, *counter-intuitive* data augmentations that simulate real-world noise by *breaking* logical consistency can be far more effective than traditional, "logically-correct" approaches.

REFERENCE

Matetić, I., Štajduhar, I., Wolf, I. and Ljubić, S. (2022) 'A Review of Data-Driven Approaches and Techniques for Fault Detection and Diagnosis in HVAC Systems', Sensors, 23(1), Art. no. 1. doi: 10.3390/s23010001.

Aghili, S. A., Rezaei, A. H. M., Tafazzoli, M., Khanzadi, M. and Rahbar, M. (2025) 'Artificial Intelligence Approaches to Energy Management in HVAC Systems: A Systematic Review', Buildings, 15(7), Art. no. 1008. doi: 10.3390/buildings15071008.

Allen-Magande, H., Khazaii, J. and Esmaeili, A. (2023) 'A Literature Review of Automated Fault Detection and Diagnostics for HVAC Systems', in ASME 2023 International Mechanical Engineering Congress and Exposition (IMECE), New Orleans, LA, USA, Paper no. IMECE2023-111611. doi: 10.1115/IMECE2023-111611.

Zhang, J., Zhang, C., Lu, J. and Zhao, Y. (2025) 'Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning', Applied Energy, 377, Art. no. 124378. doi: 10.1016/j.apenergy.2024.124378.

Crowe, E. et al. (2023) 'Empirical analysis of the prevalence of HVAC faults in commercial buildings', Science and Technology for the Built Environment, 29(10), pp. 1027–1038. doi: 10.1080/23744731.2023.2263324.

Yan, K., Yang, B. and Zhuang, C. (2025) 'Generative Models for HVAC Fault Detection and Diagnosis in Indoor and Built Environment', Journal of Building Design and Environment, 3, Art. no. 202556. doi: 10.70401/jbde.2025.0012.

Abdollah, M. A. F. (2024) Data driven Fault Detection and Diagnostics for HVAC systems in buildings. Ph.D. dissertation. Milan, Italy: Departement of Energy, Politecnico di Milano.

Sharma, V. and Mistry, V. (2023) Automated Fault Detection and Diagnostics in HVAC Systems. Zenodo. doi: 10.5281/zenodo.11079964.

Singh, V., Mathur, J. and Bathia, A. (2022) 'A comprehensive review of fault detection, diagnostics, prognostics, and fault modelling in HVAC systems', International Journal of Refrigeration, 144, pp. 142–162. doi: 10.1016/j.ijrefrig.2022.08.017.

Chen, Z. et al. (2023) 'A review of data-driven fault detection and diagnostics for building HVAC systems', Applied Energy, 339, Art. no. 121030. doi: 10.1016/j.apenergy.2023.121030.

Li, G. et al. (2021) 'Review on Fault Detection and Diagnosis Feature Engineering in Building Heating, Ventilation, Air Conditioning and Refrigeration Systems', IEEE Access, 9, pp. 2153-2187. doi: 10.1109/ACCESS.2020.3040980.

Kašiković, M. and Anđelković, A. (2025) 'Fault detection and diagnosis methods in HVAC systems', in 55th International HVAC&R Congress and Exhibition. doi: 10.24094/kghk.024.1.075.

Wu, R., Ren, Y., Tan, M. and Nie, L. (2024) 'Fault diagnosis of HVAC system with imbalanced data using multi-scale convolution composite neural network', Building Simulation, 17(3), pp. 1–16. doi: 10.1007/s12273-023-1086-1.

Gao, X. et al. (2025) 'Class-imbalanced domain generalization fault diagnosis for chiller based on simulation data', Science and Technology for the Built Environment, pp. 1–14. doi: 10.1080/23744731.2025.2556955.

Chen, H. et al. (2024) 'Review of imbalanced fault diagnosis technology based on generative adversarial networks', Journal of Computational Design and Engineering, 11(5), pp. 99–124. doi: https://doi.org/10.1093/jcde/qwae075.

Granderson, J. et al. (2023) 'A labeled dataset for building HVAC systems operating in faulted and fault-free states', Scientific Data, 10, Art. no. 342. doi: https://doi.org/10.1038/s41597-023-02197-w.

Sol, J. et al. (2024) 'Sim-to-Real Domain Adaptation for Deformation Classification', in 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, pp. 2225-2231. doi: 10.1109/SMC54092.2024.10831103.

Mikołajewska, E. et al. (2025) 'Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0', Appl. Sci., 15(6), Art. no. 3166. doi: 10.3390/app15063166.

Deng, Z. et al. (2025) 'Research on equipment fault diagnosis model based on gan and inverse PINN: Solutions for data imbalance and rare faults', PLOS ONE, 20(5), Art. no. e0324180. doi: 10.1371/journal.pone.0324180.

Forsberg, F. (2021) Domain Adaptation to meet the Reality-Gap from Simulation to Reality. Master of Science Thesis. Linköping, Sweden: Department of Electrical Engineering, Linköping University.

Ugurlu, H. I., Pham, X. H. and Kayacan, E. (2022) 'Sim-to-Real Deep Reinforcement Learning for Safe End-to-End Planning of Aerial Robots', Robotics, 11(5), Art. no. 109. doi: 10.3390/robotics11050109.

Rizzardo, C., Chen, F. and Caldwell, D. (2023) 'Sim-to-real via latent prediction: Transferring visual non-prehensile manipulation policies', Front. Robot. AI, 9, Art. no. 1067502. doi: 10.3389/frobt.2022.1067502.

Chen, R. J. et al. (2021) 'Synthetic data in machine learning for medicine and healthcare', Nat. Biomed. Eng., 5(6), pp. 493–497. doi: 10.1038/s41551-021-00751-8.

Alberts, M. et al. (2024) 'Transitioning from Simulation to Reality: Applying Chatter Detection Models to Real-World Machining Data', Machines, 12(12), Art. no. 923. doi: 10.3390/machines12120923.

Gordienko, Y. et al. (2025) 'Effect of natural and synthetic noise data augmentation on physical action classification by brain– computer interface and deep learning', Front. Neuroinform., 19, Art. no. 1521805. doi: 10.3389/fninf.2025.1521805.

Joshi, S. et al. (2025) 'Towards Mitigating Spurious Correlations in the Wild: A Benchmark and a more Realistic Dataset', arXiv preprint arXiv:2306.11957. doi: 10.48550/arXiv.2306.11957.