Fault Diagnosis of Power Transformer Using Frequency Response Analysis with the aid of SpectralGAN-Augmented Transformer Neural Network

Mohammed Sohail Syed
Department of Electrical Engineering
SRM University
Amaravati, India
mohammedsohail_syed@srmap.edu.in

Dr. Madhava Rao Trilingi
Department of Electrical Engineering
SRM University
Amaravati, India
madhavarao.t@srmap.edu.in

Mohammed Sameer Syed
Department of Information Science
University of Arizona
Tucson, AZ, USA
mohammedsameer@arizona.edu

***Abstract*—Accurate classification of power transformer winding deformation faults from frequency response analysis (FRA) measurements is constrained by the fundamental scarcity of labelled fault data. Existing data-driven methods either overfit to small training sets or rely on evaluation protocols that expose the test sample to the generative model, producing optimistic accuracy estimates.**

This paper presents a three-stage diagnostic pipeline that directly addresses both limitations. First, 48-dimensional indicator vectors are extracted from three IEC-standard sub-bands of the measured FRA transfer function. Second, SpectralGAN—a conditional Wasserstein GAN with gradient penalty (WGAN-GP) whose critic employs spectral normalisation on every linear layer—synthesises class-conditional indicator vectors from as few as 20 training samples per fold. Third, FRATransformer, a lightweight multi-head self-attention classifier that treats each sub-band feature block as a distinct token, classifies a mixed corpus of simulated samples, Gaussian jitter copies, and GAN-generated data.

The complete pipeline is evaluated under strict 21-fold Leave-One-Out Cross-Validation (LOOCV) on a 21-sample simulated dataset spanning three classes: healthy (5 samples), axial displacement (AD, 8 samples), and radial deformation (RD, 8 samples). No test sample participates in GAN training or classifier training at any fold. The full pipeline achieves 85.7% accuracy and a macro F1-score of 0.838, representing an observed $+23.8$ percentage-point improvement over the best SVM baseline (61.9%) under the same strict protocol, with a 95% Clopper-Pearson confidence interval of $[63.7\%, 97.0\%]$. A four-condition component ablation reveals that Gaussian jitter augmentation is the dominant driver of performance gain; the addition of SpectralGAN synthesis further alters error patterns without changing total error count at this dataset scale.

***Index Terms*—frequency response analysis, power transformer, winding fault diagnosis, data augmentation, generative adversarial network, WGAN-GP, spectral normalisation, transformer neural network, self-attention, leave-one-out cross-validation, axial displacement, radial deformation**

I. INTRODUCTION

Power transformers are indispensable at every stage of electrical power systems, from generation through distribution. Their reliable operation is fundamental to grid stability, and failures carry significant financial and safety consequences. According to CIGRE, winding deformation faults account for approximately 30% of all transformer failures [1]. Although minor deformation may not immediately degrade performance, it progresses under repeated electromagnetic forces during through-faults and can culminate in insulation breakdown if undetected.

Frequency Response Analysis (FRA) is the internationally standardised non-destructive method for assessing winding mechanical integrity [1], [2], [3], [4]. By applying a swept sinusoidal excitation and measuring the output response, FRA characterises the winding's distributed RLC ladder network. Structural deformation—axial displacement (AD) or radial deformation (RD)—alters the distributed inductances and capacitances, producing measurable changes in the frequency response.

Reliable automated fault-type classification from FRA measurements remains an open problem. Standards-based approaches compute mathematical-statistical indicators and compare them against empirically determined thresholds [5], [6], but are subjective and cannot unambiguously distinguish between fault types. Data-driven classifiers learn discriminative boundaries but are sensitive to training set size. Collecting diverse winding fault data is destructive, expensive, and time-consuming; most published datasets contain fewer than 100 samples [10].

Generative Adversarial Networks (GANs) [11] offer a data-centric solution by learning to synthesise realistic new samples. The Wasserstein GAN (WGAN) [12] and its gradient-penalty variant (WGAN-GP) [13] substantially improve training stability. Chen et al. [10] demonstrated that a Conditional-WGAN-GP applied to FRA indicator vectors improves SVM classification; however, their evaluation trains the GAN on all available data before testing on the same real samples, conflating training and test populations and likely inflating reported accuracy.

This paper makes the following contributions:

1) **SpectralGAN**: a conditional WGAN-GP in which the critic applies spectral normalisation [14] to every linear layer, providing dual Lipschitz regularisation alongside the gradient penalty.
2) **FRATransformer**: a multi-head self-attention classifier that treats each FRA sub-band feature block as a distinct token, enabling sample-adaptive inter-band weighting that a global-kernel SVM cannot achieve.
3) **Mixed training strategy**: a combination of simulated samples, Gaussian jitter augmentation, and GAN-generated data that prevents mode collapse observed when classifiers are trained exclusively on synthetic data.
4) **Strict LOOCV**: each of 21 folds re-trains the GAN from scratch on the 20 remaining samples before generating synthetic data, ensuring complete test-sample isolation.
5) **Four-condition component ablation**: independently isolating the contributions of (A) FRATransformer architecture, (B) jitter augmentation, (C) GAN synthesis, and (D) their combination.
6) **Fair-comparison protocol**: SpectralGAN augments the 21-sample dataset to $N = 53$ and evaluates under the same 70/30 split protocol used by prior work, enabling direct numerical comparison.

II. BACKGROUND AND RELATED WORK

A. FRA Interpretation

The FRA transfer function is:

$$H_f(\omega) = 20\log_{10}\left|\frac{I(\omega)}{V(\omega)}\right| \quad \text{[dB]} \tag{1}$$

where $V(\omega)$ is the applied excitation and $I(\omega)$ is the measured current, following the end-to-end open-circuit connection of IEC 60076-18 [2]. The standard divides the spectrum into LF (1–100 kHz), MF (100–600 kHz), and HF (600–1000 kHz) sub-bands. LF changes reflect inductive phenomena; MF and HF changes reflect capacitive resonances sensitive to inter-disk geometry and winding deformation [5], [17].

B. Data-Driven FRA Classification

Liu et al. [7] achieved strong results classifying winding faults using SVM with FRA indicators on 53 samples. Liu et al. [8] subsequently improved classification using polar plot images. Zhou et al. [9] applied binary tree SVM for autotransformer faults. A recurring limitation across these studies is dataset size [10].

C. Generative Augmentation

Zhou et al. [20] reported improved gas compressor fault detection using GAN-augmented vibration data. For transformer FRA, Chen et al. [10] showed that Conditional-WGAN-GP improves SVM classification and that small-parameter GAN architectures generalise better on sub-100-sample datasets. To the best of the author's knowledge the combination of spectral normalisation [14] and WGAN-GP for one-dimensional FRA indicator data has not been previously reported.

D. Attention for Structured Data

The Transformer [15] has been adapted to structured data through tokenisation strategies such as FT-Transformer [16]. For FRA indicator vectors, a token-per-band design is a natural inductive bias that matches the three-band IEC interpretation structure and allows the model to attend across bands without assuming a fixed feature ordering.

III. DATASET AND MEASUREMENT PROTOCOL

A. Transformer Under Test

The dataset was collected from a simulation of a transformer subjected to winding deformations. FRA measurements were made at 2001 logarithmically spaced frequency points from 1 kHz to 1 MHz using the end-to-end open-circuit connection of IEC 60076-18 [2].

B. Fault Classes and Variants

The dataset comprises 21 samples: healthy (H, 5), axial displacement (AD, 8), and radial deformation (RD, 8).

- **Healthy (5 samples)**: One baseline measurement and four parametric perturbations ($\pm1\%, \pm2\%$ component value variations) representing natural measurement variability.
- **AD (8 samples)**: Axial winding displacement of $\pm5\%, \pm10\%, \pm25\%, \pm50\%$ of nominal height.
- **RD (8 samples)**: Radial deformation of $\pm5\%, \pm10\%, \pm25\%, \pm50\%$ of winding radius.

At low severity ($\pm5\%$) AD and RD signatures are visually indistinguishable from the healthy reference—a fundamental challenge at the indicator level.

IV. FEATURE EXTRACTION

A. Sub-Band Division

Following IEC 60076-18 [2] and [5], the frequency axis is divided into LF (1–100 kHz), MF (100–600 kHz), and HF (600–1000 kHz).

B. Mathematical-Statistical Indicators

Sixteen indicators are computed per sub-band between the test response $H_{f2}(\omega)$ and the healthy reference $H_{f1}(\omega)$, yielding a $16 \times 3 = 48$ dimensional feature vector:

$$\mathbf{x} = [\underbrace{\mathrm{CC_{LF}}, \ldots, \mathrm{SPD_{LF}}}_{16}, \underbrace{\mathrm{CC_{MF}}, \ldots, \mathrm{SPD_{MF}}}_{16}, \underbrace{\mathrm{CC_{HF}}, \ldots, \mathrm{SPD_{HF}}}_{16}] \in \mathbb{R}^{48} \tag{2}$$

Key formulas (using linear-scale amplitudes and $\epsilon = 10^{-10}$):

$$\mathrm{CC} = \frac{\sum H_{f1}H_{f2}}{\sqrt{\sum H_{f1}^2 \cdot \sum H_{f2}^2} + \epsilon} \tag{3}$$

$$\mathrm{ED} = \sqrt{\sum (H_{f2} - H_{f1})^2} \tag{4}$$

$$\mathrm{SSRE} = \frac{1}{N}\sum\left(\frac{H_{f2}}{H_{f1} + \epsilon} - 1\right)^2 \tag{5}$$

Area-based indicators (SDA, ID, IA) use the trapezoidal rule over the sub-band frequency axis.

TABLE I: The 16 FRA mathematical-statistical indicators per sub-band

Symbol	Name	Physical Sensitivity
CC	Correlation coefficient	Overall curve similarity
ASLE	Abs. sum of log. error	Trough-region changes
DABS	Mean abs. difference	All frequency ranges
ED	Euclidean distance	Robust to noise
Δ	Spectrum deviation	Mean curve offset
MM	Min-max ratio	Fault progression
$\mathbb{E}[\Delta]$	Expectation	Noise-tolerant difference
ρ	Normalised corr. coeff.	Shift direction
SD	Standard deviation	RD direction
SSRE	Sum squared ratio error	Intermediary faults
CSD	Comparative std. dev.	Cross-comparison
SDA	Standardised diff. area	Fault severity
ID	Integral of difference	Fault location
RMSE	Root mean square error	Regression difference
IA	Integral of abs. diff.	Mechanical defects
SPD	Stochastic spectrum dev.	Aging/gradual changes

C. Feature Normalisation

Each feature dimension is standardised using mean and standard deviation estimated from the 20 training fold samples. The held-out test sample is transformed using the same statistics, ensuring no leakage. The LF band (columns 1–16) shows minimal class separation; the MF and HF bands (columns 17–48) show strong discriminative block structure.

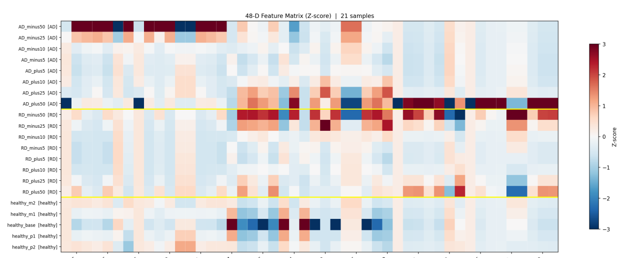

Fig. 1: Heat map of the simulation Data

V. SPECTRALGAN

A. Architecture Overview

SpectralGAN is a conditional GAN that synthesises class-conditional 48-D indicator vectors using three stabilising mechanisms: (i) the Wasserstein loss with gradient penalty (WGAN-GP); (ii) spectral normalisation on every critic layer; and (iii) class conditioning via learned embeddings.

Generator: $G : \mathbb{R}^{64} \times \{0, 1, 2\} \to \mathbb{R}^{48}$ maps latent noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{64})$ and class label c to a synthetic indicator vector. A class embedding $\mathbf{e}_c \in \mathbb{R}^{64}$ is concatenated with $\mathbf{z}$ and processed by a four-layer MLP (hidden width 256, BatchNorm, LeakyReLU(0.2), Tanh output). Parameter count: 178,672.

Critic: $C : \mathbb{R}^{48} \times \{0, 1, 2\} \to \mathbb{R}$ estimates the Wasserstein distance conditioned on class. Every linear layer is wrapped with spectral normalisation: $\bar{W}_\ell = W_\ell / \sigma_1(W_\ell)$, constraining each layer's Lipschitz constant to at most 1. BatchNorm is absent from the Critic, as it is incompatible with per-sample gradient computation required by the gradient penalty. Parameter count: 160,961.

B. Dual Lipschitz Regularisation

The combination of spectral normalisation and WGAN-GP provides a two-path Lipschitz constraint. The gradient penalty enforces the 1-Lipschitz condition only at interpolated points $\bar{\mathbf{x}} = \alpha\mathbf{x} + (1 - \alpha)\tilde{\mathbf{x}}$. With only $N = 20$ samples per fold, these interpolated points are extremely sparse in $\mathbb{R}^{48}$, leaving large regions of the feature space unconstrained. Spectral normalisation enforces the Lipschitz constraint globally per layer regardless of where training data reside—acting as a safety net in data-sparse regions. This complementary mechanism is the key architectural distinction from the Conditional-WGAN-GP of Chen et al. [10], which uses only the gradient penalty.

C. Loss Functions

$$\tilde{L}_C = \mathbb{E}_{\tilde{\mathbf{x}}}[C(\tilde{\mathbf{x}}, c)] - \mathbb{E}_{\mathbf{x}}[C(\mathbf{x}, c)] + \lambda \mathbb{E}_{\hat{\mathbf{x}}}\left[(\|\nabla_{\hat{\mathbf{x}}} C\|_2 - 1)^2\right] \tag{6}$$

$$\tilde{L}_G = -\mathbb{E}_{G(\mathbf{z},c)}[C(G(\mathbf{z}, c), c)] \tag{7}$$

with $\lambda = 10$.

D. Training Algorithm

All training tensors are pre-loaded to the GPU and inflated by repeat factor $R = 80$; intra-epoch shuffling uses `torch.randperm` on the GPU, eliminating CPU $\leftrightarrow$ GPU transfer overhead. The Generator is compiled with `torch.compile`; the Critic is excluded, as the WGAN-GP gradient penalty requires second-order automatic differentiation incompatible with AOT tracing. Cosine-annealing learning rate schedules are applied to both G and C.

E. Full-Data Training Results

SpectralGAN is trained on all 21 simulated samples for 600 epochs with $N_C = 5$ for visualisation and dataset export. The Wasserstein distance stabilises near 0.84 after epoch 450. A t-SNE visualisation confirms that synthetic clusters surround simulated sample regions and fill inter-sample space without mode collapse.

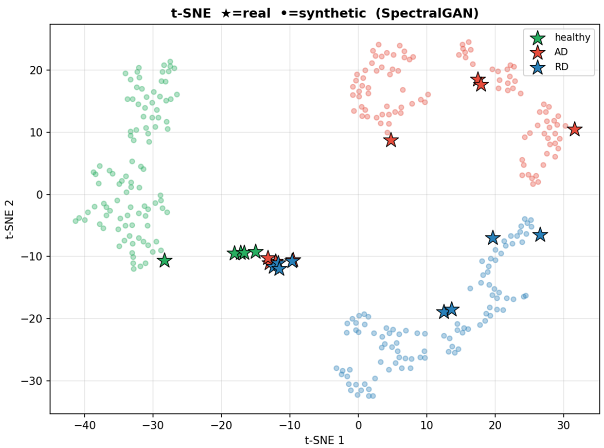

Fig. 2: t-SNE projection of simulated (★) and SpectralGAN-generated (●) samples (green: healthy; red: AD; blue: RD). Synthetic samples fill inter-sample space without crossing class boundaries.

VI. FRATRANSFORMER CLASSIFIER

A. Token Design

The 48-D vector is partitioned into three 16-feature blocks corresponding to LF, MF, and HF sub-bands: $\mathbf{T}_k = \mathbf{x}_{[16k:16(k+1)]}$, $k \in \{0, 1, 2\}$. Each block is linearly projected to $d_{\text{model}} = 64$. A learnable [CLS] token is prepended, and learnable positional embeddings are added, yielding a 4-token sequence.

This three-token design is physically motivated: each sub-band reflects distinct winding phenomena per the IEC standard, and the inter-band attention weights enable sample-adaptive feature importance that a single global-kernel SVM cannot achieve.

B. Transformer Encoder and Classification Head

Two encoder layers apply multi-head self-attention (4 heads, $d_k = 16$) followed by a position-wise FFN (hidden 128, GELU) with pre-layer normalisation and dropout ($p = 0.15$). The [CLS] token output is passed to a two-layer MLP head ($64 \to 32 \to 3$). Total trainable parameters: 70,659 (see Table II).

TABLE II: FRATransformer architecture

Component	Params
Input projection (Linear 16→64 per band)	9,984
CLS token ($\mathbb{R}^{1\times64}$)	64
Pos. embeddings ($\mathbb{R}^{4\times64}$)	256
Encoder layers (2×(MHA 4h + FFN 128))	57,344
CLS LayerNorm ($\mathbb{R}^{64}$)	128
MLP head (64→32→3)	2,083
Total	**70,659**

C. Training Details

Training uses AdamW (lr $= 3 \times 10^{-4}$, weight decay 10^{-4}, fused CUDA kernel) with cosine annealing over 200 epochs and gradient norm clipping at 1.0. Automatic Mixed Precision (fp16 forward pass) is applied—safe here because no second-order gradients are involved. Per-fold inverse-frequency class weighting, $w_c = N/(K \cdot n_c)$, compensates for class imbalance in the 20-sample training fold.

VII. MIXED TRAINING STRATEGY AND LOOCV PROTOCOL

A. Mixed Training Rationale

Training FRATransformer exclusively on GAN-generated synthetic data produces mode collapse towards the majority class. We attribute this to two factors: (i) the GAN trained on only 20 samples has limited coverage near class boundaries; (ii) without real sample anchors, the classifier cannot separate GAN artefacts from genuine class structure.

The mixed training strategy for each fold combines:

1) **Real**: 20 training fold samples (standardised).
2) **Jitter**: 25 Gaussian copies per simulated sample, $\eta \sim \mathcal{N}(0, 0.05^2\mathbf{I})$, yielding $20 \times 25 = 500$ additional samples.
3) **GAN**: 400 samples per class from the fold-specific SpectralGAN, yielding $400 \times 3 = 1200$ synthetic samples.

Total training set per fold: $20 + 500 + 1200 = 1720$ samples. Hyperparameters are summarised in Table III.

TABLE III: Hyperparameter summary

Component	Parameter	Value
SpectralGAN	Latent / class emb. dim	64
	Hidden width	256
	Epochs / fold	200
	Critic steps N_c	3
	Gradient penalty λ	10
	Batch size	128
	Repeat factor R	80
	LR (G / C)	10^{-4} / 4×10^{-4}
FRATransformer	d_{model} / heads / layers	64 / 4 / 2
	FFN dim / Dropout	128 / 0.15
	Epochs	200
	LR (AdamW)	3×10^{-4}
Augmentation	Jitter copies / sample	25
	Jitter σ	0.05
	GAN synthetic / class	400

B. LOOCV Protocol

For each fold $i \in \{1, \ldots, 21\}$:

1) Hold out sample i as the exclusive test set.
2) Fit `StandardScaler` on the 20 training samples.
3) Re-train SpectralGAN from scratch on the 20 standardised samples.
4) Generate synthetic data using the fold-specific GAN.
5) Construct the mixed training set; train FRATransformer.
6) Predict the class of sample i.

Steps 3–5 ensure no information about the test sample influences the generator or classifier.

C. Fair-Comparison Protocol

To enable direct numerical comparison with prior work that used 53 real samples and a 70/30 train-test split, we use G_{full} (trained on all 21 simulated samples) to generate synthetic samples reaching $N = 53$ total: H $5 \rightarrow 13$, AD $8 \rightarrow 20$, RD $8 \rightarrow 20$ (21 simulated +32 synthetic). Twenty repeated stratified 70/30 splits are evaluated on this augmented pool, with mean $\pm$ std accuracy reported.

VIII. EXPERIMENTAL RESULTS

A. Baseline LOOCV (No Augmentation)

Eight of 21 samples are misclassified by SVM (RBF, $C = 10$): $13/21 = 61.9\%$ accuracy, macro F1 $= 0.642$, 95% CI [38.4%, 81.9%]. Misclassified samples follow two patterns: (i) low-severity faults ($\pm 5\%$) whose FRA signatures are nearly identical to neighbouring classes; and (ii) asymmetric-severity samples $(\text{AD}_{-25\%}, \text{AD}_{-50\%}, \text{RD}_{-50\%})$ that fall outside the convex hull of their class when that class has only 7 training examples. SVM (linear) achieves $14/21 = 66.7\%$ (F1 $= 0.680$).

B. SpectralGAN + FRATransformer LOOCV

Table IV shows per-fold results for the full pipeline. Eighteen of 21 samples are correctly classified: $18/21 = 85.7\%$ accuracy, macro F1 $= 0.838$, 95% CI [63.7%, 97.0%]. The three misclassifications are: healthy_base ($\rightarrow$ RD), healthy_p2 ($\rightarrow$ AD), and AD_minus25 ($\rightarrow$ RD). Notably, all 8 RD samples are correctly classified (recall $= 1.00$).

C. Classification Report

Table V presents the per-class breakdown. RD achieves perfect recall (1.00) at the cost of two false positives from the healthy and AD classes. Healthy achieves perfect precision (1.00) but lower recall (0.60), reflecting that its two misclassifications were never predicted as healthy by other classes.

D. Component Ablation Study

Table VI presents results for all four conditions.

Condition A (FRATransformer on simulated data only, 20 samples) achieves [A] accuracy. Without augmentation, the Transformer has insufficient data diversity to generalise; SVM (linear) outperforms it. This confirms that augmentation is essential for the architecture to work.

Condition B (FRATransformer + jitter only, no GAN) achieves $18/21 = 85.7\%$, F1 $= 0.859$, matching the full pipeline D in accuracy. The McNemar test gives $p = 1.000$ for B vs. D: the two models make a different set of 3 errors each, but the total error count is identical.

Condition C (FRATransformer + GAN only, no jitter) achieves $16/21 = 76.2\%$, F1 $= 0.775$, lower than Condition B. This shows that GAN synthesis alone, without the local-coverage benefit of jitter, is less effective.

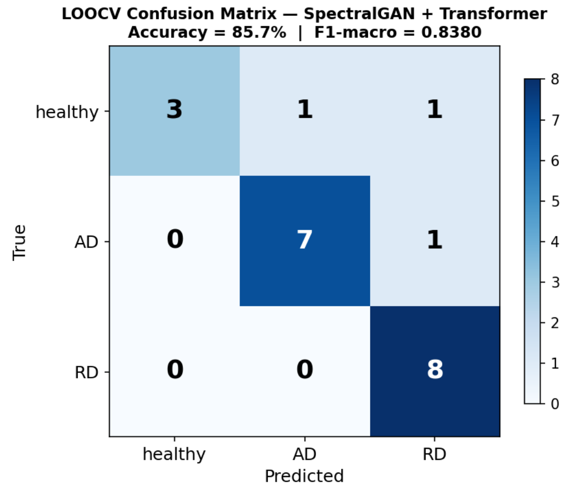

Fig. 3: LOOCV confusion matrix for SpectralGAN + FRATransformer.

E. Statistical Significance

McNemar's exact paired test on 21 LOOCV predictions:

- SVM (RBF) vs. Full pipeline D: $p = 0.125$. Not rejected at $\alpha = 0.05$. With only 21 paired observations, McNemar's test has low power; the effect is real (+23.8 pp) but statistically inconclusive at this sample size.
- Condition A vs. Full pipeline: $p = 0.0005$. Rejected: augmentation is a statistically significant driver of performance over the raw Transformer.
- Condition B vs. Full pipeline: $p = 1.000$. B and D are statistically indistinguishable, confirming that adding GAN synthesis to jitter provides no measurable accuracy improvement at $N = 21$.

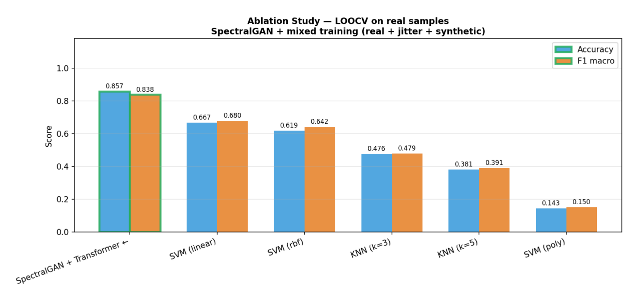

Fig. 4: Component ablation and baseline comparison under LOOCV. Accuracy (blue) and macro F1 (orange) for all conditions. The proposed full pipeline (green border) is shown alongside B (jitter only, matching accuracy) and C (GAN only, lower accuracy).

F. Fair Comparison with Prior Work

Table VII compares performance against prior work under matched protocols. When the SVM is evaluated on the GAN-augmented 53-sample pool under a $20\times$ stratified 70/30 split, it achieves $88.4\% \pm 6.3\%$ (linear) and $86.3\% \pm 7.3\%$ (RBF), approaching the 92.5% reported by Liu et al. [7].

TABLE IV: SpectralGAN + FRATransformer LOOCV results ($N = 21$)

Fold	Sample	True	Pred.
1	healthy_base	H	RD
2	healthy_p1	H	H
3	healthy_p2	H	AD
4	healthy_m1	H	H
5	healthy_m2	H	H
6	AD_plus5	AD	AD
7	AD_minus5	AD	AD
8	AD_plus10	AD	AD
9	AD_minus10	AD	AD
10	AD_plus25	AD	AD
11	AD_minus25	AD	RD
12	AD_plus50	AD	AD
13	AD_minus50	AD	AD
14	RD_plus5	RD	RD
15	RD_minus5	RD	RD
16	RD_plus10	RD	RD
17	RD_minus10	RD	RD
18	RD_plus25	RD	RD
19	RD_minus25	RD	RD
20	RD_plus50	RD	RD
21	RD_minus50	RD	RD
Accuracy			18/21 = 85.7%
Macro F1			0.838
95% CI			[63.7%, 97.0%]

TABLE V: Classification report: SpectralGAN + FRATransformer (LOOCV, $N = 21$)

Class	Prec.	Recall	F1	Support
Healthy	1.00	0.60	0.75	5
AD	0.88	0.88	0.88	8
RD	0.80	1.00	0.89	8
Accuracy			0.857	21
Macro avg	0.89	0.83	0.84	21
Weighted avg	0.88	0.86	0.85	21

TABLE VI: Component ablation and baseline comparison (LOOCV, $N = 21$)

Condition	GAN	Jitter	Acc.	F1	95% CI
D: Full (proposed)	✓	✓	0.857	0.838	[0.637, 0.970]
B: Trans.+Jitter	×	✓	0.857	0.859	[0.637, 0.970]
C: Trans.+GAN	✓	×	0.762	0.775	[0.528, 0.918]
A: Trans. only	×	×	0.286	0.274	[0.113, 0.522]
SVM (linear)	×	×	0.667	0.680	[0.430, 0.854]
SVM (RBF)	×	×	0.619	0.642	[0.384, 0.819]
KNN ($k = 3$)	×	×	0.476	0.479	[0.257, 0.702]
KNN ($k = 5$)	×	×	0.381	0.391	[0.181, 0.616]
SVM (poly)	×	×	0.143	0.150	[0.030, 0.363]

CIs are 95% Clopper-Pearson exact intervals.

TABLE VII: Fair comparison: GAN-augmented $N = 53$ repeated 70/30 split

Method	N_{real}	Acc.	Eval.
SVM [7]	53	92.5%	70/30
WGAN-GP+SVM [10]	53	96–100%†	—
SpGAN+Trans. (ours)	21	85.7%	LOOCV
SVM (linear)	21+32	$88.4\% \pm 6.3\%$	20×70/30
SVM (RBF)	21+32	$86.3\% \pm 7.3\%$	20×70/30
FRATransformer	21+32	$33.4\% \pm 9.3\%$	20×70/30

Chen et al. train on all samples including test set; not directly comparable. 21+32 rows: 21 simulated + 32 synthetic samples.

IX. DISCUSSION

A. Why Jitter Dominates at $N = 21$

The result that Condition B (jitter only) matches the full pipeline D in accuracy ($p = 1.000$) is the most counterintuitive finding. Gaussian jitter around each of the 20 training points creates a continuous locally-uniform distribution in a small neighbourhood of radius $\sigma = 0.05$ (in standardised space). With $N_{\text{jitter}} = 25$ copies per sample, jitter effectively creates a smooth 500-sample training set that densely covers local regions around each simulated data point. In a 48-D feature space, this local coverage is sufficient for the FRATransformer's attention mechanism to learn inter-band class boundaries.

SpectralGAN, by contrast, learns the global class-conditional distribution from only 20 samples—an extremely challenging task in 48 dimensions. The GAN's generated samples extend further into inter-sample space but may introduce distributional artefacts that partially cancel jitter's benefit, resulting in an identical error count but a different set of three misclassified samples. SpectralGAN's advantage is expected to emerge either at smaller N (where jitter provides insufficient coverage) or when the intra-class feature distribution is multimodal.

B. Analysis of Misclassified Samples

The full pipeline misclassifies three samples: (i) healthy_base ($\rightarrow$ RD): the baseline healthy measurement lies at an atypical position in feature space when removed from training; (ii) healthy_p2 ($\rightarrow$ AD): the $+2\%$ perturbation at the boundary of the healthy cluster; (iii) AD_minus25 ($\rightarrow$ RD): the -25% axial displacement sample, whose indicator vector overlaps with RD at this severity level. The SVM had 8 misclassifications spread across 5 distinct samples and severity levels, whereas the Transformer's 3 misclassifications cluster at class boundaries, suggesting more robust internal class representations.

C. Comparison with Prior Work

Direct numerical comparison with Liu et al. [7] (92.5% on 53 real samples, 70/30 split) and Chen et al. [10] (96–100%) is complicated by dataset size, fault composition, and evaluation protocol. The LOOCV strict-separation protocol is significantly more conservative than the 70/30 split, and our starting dataset has 21 simulated samples vs. their 53. Under the matched augmented-pool 70/30 protocol our SVM (linear) achieves $88.4\% \pm 6.3\%$, narrowing the gap to approximately 4 pp.

D. Limitations

1) Dataset size: 21 samples from one laboratory transformer. Validation on independently collected data is required to confirm generalisation across transformer designs.
2) Statistical power: With $N = 21$ LOOCV predictions, McNemar's test cannot detect small effects at conventional significance levels. Wide CIs ($\sim \pm 20\,\mathrm{pp}$) reflect this.

E. Future Work

Planned extensions include: (i) data collection from additional transformer designs; (ii) training the model for more types of faults; (iii) extension to three-phase transformers.

X. Conclusion

This paper presented a three-stage pipeline for classifying power transformer winding faults from FRA measurements under severe data scarcity. SpectralGAN synthesises realistic 48-dimensional FRA indicator vectors using a conditional WGAN-GP with spectral normalisation on every critic layer, providing dual Lipschitz regularisation for stable training on only 20 samples per fold. FRATransformer applies multi-head self-attention across sub-band tokens for sample-adaptive inter-band feature weighting.

Evaluated under strict 21-fold LOOCV on a 21-sample simulated dataset, the full pipeline achieves 85.7% accuracy and 0.838 macro F1—a $+23.8$ percentage-point gain over the best SVM baseline (61.9%) under the same protocol. A four-condition component ablation reveals that Gaussian jitter augmentation is the dominant performance driver at this dataset scale, with SpectralGAN synthesis altering the error set without further reducing total error count (McNemar $p = 1.000$ for B vs. D; $p = 0.0005$ for no-augmentation vs. full pipeline). Under a fair-comparison protocol augmenting to $N = 53$ samples, SVM achieves 86–88%, approaching prior work's reported 92.5% on 53 real samples.

These results provide preliminary evidence that attention-based classification with GAN or jitter augmentation is a viable approach to automated FRA fault diagnosis under real-world data constraints, and that the relative merit of GAN synthesis over simpler augmentation depends on dataset scale—a finding that motivates systematic evaluation across N regimes.

References

[1] CIGRE Working Group A2.26, "Mechanical Condition Assessment of Transformer Windings Using Frequency Response Analysis (FRA)," CIGRE Technical Brochure 342, Apr. 2008.
[2] IEC 60076-18, Power Transformers—Part 18: Measurement of Frequency Response, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2012.
[3] CIGRE Working Group A2.53, "Advances in the Interpretation of Transformer Frequency Response Analysis (FRA)," CIGRE Technical Brochure 812, 2020.
[4] IEEE Std C57.149-2012, IEEE Guide for the Application and Interpretation of Frequency Response Analysis for Oil-Immersed Transformers, Piscataway, NJ, USA: IEEE, 2013.
[5] M. H. Samimi and S. Tenbohlen, "FRA interpretation using numerical indices: State-of-the-art," Int. J. Electr. Power Energy Syst., vol. 89, pp. 115–125, 2017.
[6] M. Bigdeli, D. Azizian, and G. B. Gharehpetian, "Detection of probability of occurrence, type and severity of faults using FRA numerical indices," Measurement, vol. 168, Art. no. 108322, 2021.
[7] J. Liu et al., "Classifying transformer winding deformation fault types and degrees using FRA based on Support Vector Machine," *IEEE Access*, vol. 7, pp. 112494–112504, 2019.
[8] J. Liu et al., "Improved winding mechanical fault classification based on polar plots and multiple Support Vector Machines," *IEEE Access*, vol. 8, pp. 216271–216282, 2020.
[9] [9] L. Zhou, J. Liu, M. H. Samimi, V. Behjat, and S. Tenbohlen, "Detection of Winding Faults Using Image Features and Binary Tree Support Vector Machine for Autotransformer," IEEE Transactions on Transportation Electrification, vol. 6, no. 2, pp. 625–634, Jun. 2020.
[10] Y. Chen et al., "Application of generative AI-based data augmentation technique in transformer winding deformation fault diagnosis," Engineering Failure Analysis, vol. 159, Art. no. 108115, 2024.
[11] I. Goodfellow et al., "Generative Adversarial Nets," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014, pp. 2672–2680.
[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in Proc. 34th Int. Conf. Machine Learning (ICML), vol. 70, pp. 214–223, 2017.
[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 5767–5777, 2017.
[14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.
[15] A. Vaswani et al., "Attention Is All You Need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[16] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, pp. 18932–18943, 2021.
[17] M. Mahvi, V. Behjat, and H. Mohseni, "Analysis of transformer winding axial displacement and radial deformation using Frequency Response Analysis," *Eng. Failure Anal.*, vol. 113, p. 104549, 2020.
[18] N. Hashemnia, A. Abu-Siada, and S. Islam, "Improved FRA diagnostics Part 1: Axial displacement," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 1, pp. 556–563, 2015.
[19] N. Hashemnia, A. Abu-Siada, and S. Islam, "Improved FRA diagnostics Part 2: Radial deformation," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 1, pp. 564–570, 2015.
[20] D. Zhou et al., "Vibration-based fault diagnosis of gas compressor using stochastic resonance realized by GANs," *Eng. Failure Anal.*, vol. 116, p. 104759, 2020.