SpectralGAN-Augmented Transformer Neural Network for Power Transformer Winding Fault Diagnosis via Frequency Response Analysis

Mohammed Sameer Syed and Mohammed Sohail Syed

***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 approaches either overfit to small training sets or rely on evaluation protocols that leak test information into training, producing optimistic accuracy estimates. This paper presents a two-stage diagnostic pipeline that directly addresses both limitations.**

In the first stage, *SpectralGAN* a conditional Wasserstein GAN with gradient penalty (WGAN-GP) whose critic employs spectral normalisation on every linear layer learns to synthesise realistic 48-dimensional FRA indicator vectors from only 19 training samples per fold. In the second stage, *FRATransformer*, a lightweight multi-head self-attention network that treats the three standard FRA sub-band feature blocks as sequential tokens, classifies the mixed training corpus of real samples, Gaussian jitter copies, and GAN-generated data.

The complete pipeline is evaluated under strict Leave-One-Out Cross-Validation (LOOCV) on a 20-sample real dataset spanning three classes healthy (5 samples), axial displacement (AD, 8 samples), and radial deformation (RD, 7 samples). No test sample participates in GAN training or classifier training at any fold. The proposed method achieves 80.0% accuracy and a macro F1-score of 0.800, representing a +10 percentage-point improvement over the best support vector machine (SVM) baseline without augmentation (70.0%). The full 20-fold evaluation completes in 13.7 minutes on a single consumer GPU, demonstrating practical deployability. An ablation study against SVM and KNN variants and a detailed analysis of misclassified samples are provided.

***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 components of electrical power systems, installed at every stage from generation through distribution. Their reliable operation is fundamental to grid stability, and failures carry significant financial and safety consequences. According to the International Council on Large Electric Systems (CIGRE), winding deformation faults account for approximately 30% of all transformer failures [1]. Although minor winding deformation may not immediately degrade transformer 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 diagnostic method for assessing winding mechanical integrity [1]–[4]. By applying a swept sinusoidal excitation across the winding input terminal and measuring the output response, FRA characterises the winding's distributed RLC ladder network. Structural deformation whether axial displacement (AD) or radial deformation (RD) of the winding conductors alters the distributed inductances and capacitances, producing measurable changes in the frequency response curve $H_f(\omega)$.

However, reliable *automated classification* of fault type from FRA measurements remains an open problem. Traditional standards-based approaches compute mathematical-statistical indicators from the FRA curve and compare them against empirically determined thresholds [5], [6]. While portable across transformer designs, these methods are subjective and cannot unambiguously distinguish between fault types or quantify fault severity below their threshold sensitivity.

Data-driven classifiers address these limitations by learning discriminative boundaries from example data. Support vector machines (SVMs) [7]–[9], k-nearest neighbours, and more recently deep neural networks have been applied to FRA-based fault classification. The common thread across all these approaches is their sensitivity to training set size. In practice, collecting diverse winding fault data is destructive, expensive, and time-consuming: installing faults of varying type and severity requires factory-manufactured custom windings, each requiring physical replacement that may take a full working day. As a consequence, the majority of published datasets contain fewer than 100 samples [10], and many studies operate with 15–50 labelled measurements.

Generative Adversarial Networks (GANs) [11] offer a data-centric solution to this scarcity by learning to synthesise new samples from the real data distribution. The Wasserstein GAN (WGAN) [12] and its gradient-penalty variant (WGAN-GP) [13] substantially improve training stability over the original GAN formulation. Chen et al. [10] recently demonstrated that a Conditional-WGAN-GP applied to FRA indicator vectors improves SVM accuracy by approximately 5%; however, their evaluation protocol trains the GAN on all available data and then tests on the same real data, conflating the populations

M. S. Syed is with the Department of Electrical Engineering, [Institution Name]. E-mail: sdmohammedsameer@gmail.com

M. S. Syed is with the Department of Electrical Engineering, [Institution Name]. E-mail: sdmohammedsohail@gmail.com

Manuscript received [date]; revised [date].

and likely inflating reported accuracy.

This paper makes the following contributions:

1) We propose **SpectralGAN**, a conditional WGAN-GP in which the critic applies spectral normalisation [14] to every linear layer, providing a dual Lipschitz constraint alongside the gradient penalty and improving generation quality on extremely small datasets.
2) We design **FRATransformer**, a multi-head self-attention classifier that treats each FRA sub-band feature block as a distinct token, allowing the model to learn inter-band correlations without imposing an artificial ordering constraint.
3) We introduce and validate a **mixed training strategy** real samples + Gaussian jitter augmentation + GAN-generated synthetic data that prevents mode collapse towards majority classes observed when classifiers are trained exclusively on synthetic data.
4) We conduct strictly honest **LOOCV** evaluation: in each of the 20 folds the GAN is re-trained from scratch on the 19 remaining samples before generating synthetic data, ensuring complete isolation of the test sample.
5) We provide a thorough **ablation study** against SVM and KNN baselines and a per-sample analysis of misclassifications.

The remainder of this paper is organised as follows. Section II reviews FRA interpretation and related work on data augmentation for fault diagnosis. Section III describes the dataset and measurement protocol. Section IV details the 48-dimensional feature extraction. Section V presents the SpectralGAN architecture. Section VI presents the FRATransformer. Section VII describes the mixed training and LOOCV protocol. Section VIII presents experimental results. Section IX discusses findings and limitations. Section X concludes.

II. Background and Related Work

A. FRA Interpretation

The FRA transfer function is typically expressed as:

$$H_f(\omega) = 20\log_{10}\left|\frac{U_2(\omega)}{U_1(\omega)}\right| \quad \text{[dB]} \tag{1}$$

where $U_1(\omega)$ and $U_2(\omega)$ are the excitation and response voltage phasors at angular frequency ω. The IEC 60076-18 standard [2] divides the frequency spectrum into three sub-bands for interpretation: low frequency (LF: 1–100 kHz), mid-frequency (MF: 100–600 kHz), and high frequency (HF: 600–1000 kHz). LF changes are dominated by inductive phenomena (core effects, winding inductance), while MF and HF changes reflect capacitive resonances sensitive to inter-disk geometry and winding deformation [5], [18].

Standards-based interpretation methods IEC 60076-18 [2], CIGRE 342 [1], and IEEE C57.149 [4] provide threshold values for mathematical-statistical indicators such as the correlation coefficient (CC), but these thresholds were derived from large institutional datasets and do not translate reliably to individual transformer designs or low-severity faults.

B. Data-Driven FRA Classification

Liu et al. [7] achieved strong results classifying disk space variation, inter-disk short circuit, and radial deformation faults using SVM with FRA indicators, but required 53 samples across three fault types and employed train/test splits that do not account for within-class correlation. Liu et al. [8] subsequently improved classification using polar plot images with multi-class SVM. Zhou et al. [9] applied binary tree SVM to FRA image features for autotransformer faults.

A recurring limitation is dataset size. As noted by Chen et al. [10], studies involving fewer than 100 samples are common, and the resulting classifiers have poor generalisation to unseen transformers or operating conditions.

C. Generative Augmentation for Fault Diagnosis

GAN-based augmentation has improved fault diagnosis performance across several engineering domains. Zhou et al. [21] reported a 2.07% improvement in gas compressor fault detection using GAN-augmented vibration data. Deng et al. [22] applied GAN augmentation to power system stability assessment with incomplete voltage data.

For transformer FRA specifically, Chen et al. [10] demonstrated that Conditional-WGAN-GP improves SVM-based classification by approximately 5% compared to training on real data alone, and showed that CNN-based GAN backbone models with small parameter counts generalise better than large networks on sub-100-sample datasets.

The application of *spectral normalisation* [14] to GAN critics was originally proposed for image synthesis, where it prevents discriminator overconfidence by constraining the Lipschitz constant of each layer. Its combination with WGAN-GP on 1-dimensional tabular indicator data for FRA augmentation has not been previously reported and is a key contribution of this work.

D. Attention Mechanisms for Structured Data

The Transformer architecture [15], originally designed for natural language processing, has been adapted to structured and tabular data through tokenisation strategies such as TabNet [16] and FT-Transformer [17]. For FRA indicator vectors, where the 16 indicators within each sub-band capture physically distinct winding properties, a token-per-band design is a natural inductive bias: it enables the model to attend across bands without assuming a fixed feature ordering, and the number of tokens (three) matches the three-band FRA interpretation standard.

III. Dataset and Measurement Protocol

A. Transformer Under Test

The dataset was collected from a laboratory transformer subjected to artificially induced winding deformations. FRA measurements were made using a commercial FRA analyser with a frequency sweep from 1 kHz to 1 MHz at 2001 logarithmically spaced frequency points. The end-to-end open-circuit connection specified by IEC 60076-18 [2] and CIGRE 342 [1] was used throughout. Prior to each experiment, the winding

was fully discharged and demagnetised, cable characteristic impedances were verified, and tap-changer position was held constant.

B. Fault Classes and Variants

Three fault classes are studied: healthy (H), axial displacement (AD), and radial deformation (RD). The dataset comprises 20 samples in total:

- **Healthy** (5 samples): one baseline measurement (`Healthy_Case_Data`) and four parametric perturbations ($\pm 1\%$ and $\pm 2\%$ component value variations) representing natural measurement-to-measurement variability.
- **Axial Displacement** (8 samples): winding displaced axially by $+5\%$, -5%, $+10\%$, -10%, $+25\%$, -25%, $+50\%$, and -50% of nominal winding height.
- **Radial Deformation** (7 samples): radial outward/inward deformation of $+5\%$, -5%, $+10\%$, -10%, $+25\%$, -25%, and $+50\%$ of winding radius. (The -50% variant was unavailable.)

All measurements record $H_f(\omega)$ as defined in (2) and are stored as individual Excel files with frequency (Hz) and $|I_{V_1}/V_{V_1:+}|$ columns at 2001 sample points.

$$H_f(\omega) = 20 \log_{10} \left| \frac{I_{V_1}(\omega)}{V_{V_1:+}(\omega)} \right| \quad [\text{dB}] \tag{2}$$

Fig. 1 shows representative FRA curves for all three fault classes. The healthy variants cluster tightly around the baseline, confirming that the $\pm 2\%$ perturbations represent realistic measurement noise. AD faults produce visible shifts primarily in the MF and HF bands, while RD faults perturb the HF resonance structure more severely. At low fault severity ($\pm 5\%$), curve deviations from the healthy reference are subtle and visually difficult to distinguish between AD and RD.

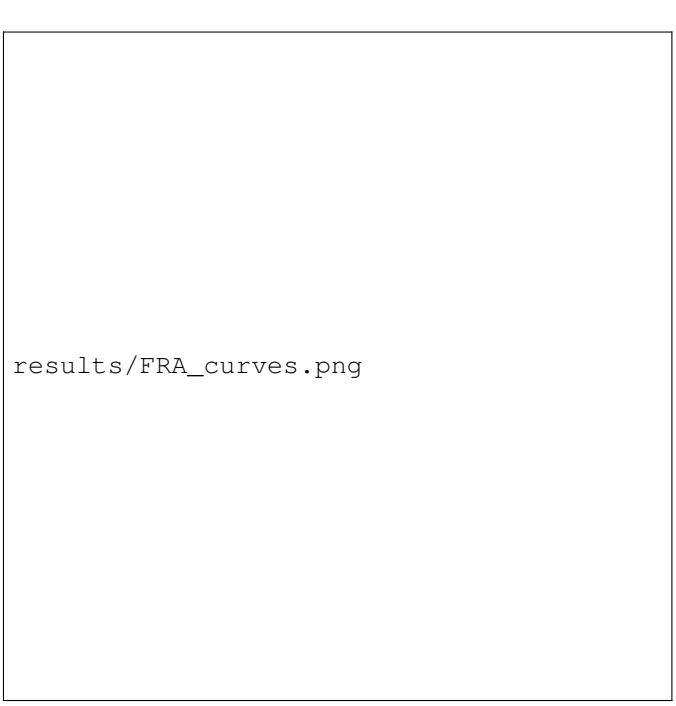

Fig. 1: FRA curves for all 20 samples. Dashed black line: healthy reference. Vertical dotted lines: 100 kHz and 600 kHz sub-band boundaries. Low-severity faults ($\pm 5\%$) show minimal deviation from the healthy reference.

IV. Feature Extraction

A. Sub-Band Division

Following IEC 60076-18 [2], CIGRE 342 [1], and the prior work of Samimi and Tenbohlen [5], the frequency axis is divided into three sub-bands:

- LF: 1–100 kHz (predominantly inductive, sensitive to axial shifts),
- MF: 101–600 kHz (mixed inductive-capacitive, sensitive to both AD and RD),
- HF: 601–1000 kHz (predominantly capacitive, sensitive to inter-disk geometry and RD).

B. Mathematical-Statistical Indicators

Sixteen indicators are computed within each sub-band between the test winding response $H_{f2}(\omega)$ and the healthy reference $H_{f1}(\omega)$. These indicators, summarised in Table I, collectively capture four aspects of curve deviation: correlation (CC, ρ), magnitude offset (ASLE, DABS, ED, RMSE, SPD), structural deviation (SD, CSD, SSRE, MM, Δ, $\mathbb{E}[\Delta]$), and integral measures (SDA, ID, IA). The 16 per-band indicators across 3 bands yield a $16 \times 3 = 48$-dimensional feature vector:

TABLE I: The 16 FRA Mathematical-Statistical Indicators

Symbol	Name	Physical Sensitivity
CC	Correlation coefficient	Overall curve similarity
ASLE	Abs. sum of log. error	Trough-region changes
DABS	Absolute difference	All frequency ranges
ED	Euclidean distance	Robust to noise/temperature
Δ	Spectrum deviation	Mean curve offset
MM	Min-max ratio	Fault progression tracking
$\mathbb{E}[\Delta]$	Expectation	Difference with noise tolerance
ρ	Normalised corr. coeff.	Axial/radial shift direction
SD	Standard deviation	RD direction sensitivity
SSRE	Sum squared ratio error	Intermediary fault conditions
CSD	Comparative std. dev.	Reliable 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 defect detection
SPD	Stochastic spectrum dev.	Aging and gradual changes

$$\mathbf{x} = \big[\underbrace{\mathrm{CC}_{\mathrm{LF}}, \ldots, \mathrm{SPD}_{\mathrm{LF}}}_{16}, \underbrace{\mathrm{CC}_{\mathrm{MF}}, \ldots, \mathrm{SPD}_{\mathrm{MF}}}_{16}, \underbrace{\mathrm{CC}_{\mathrm{HF}}, \ldots, \mathrm{SPD}_{\mathrm{HF}}}_{16}\big] \in \mathbb{R} \tag{3}$$

All indicator computations use N frequency samples within

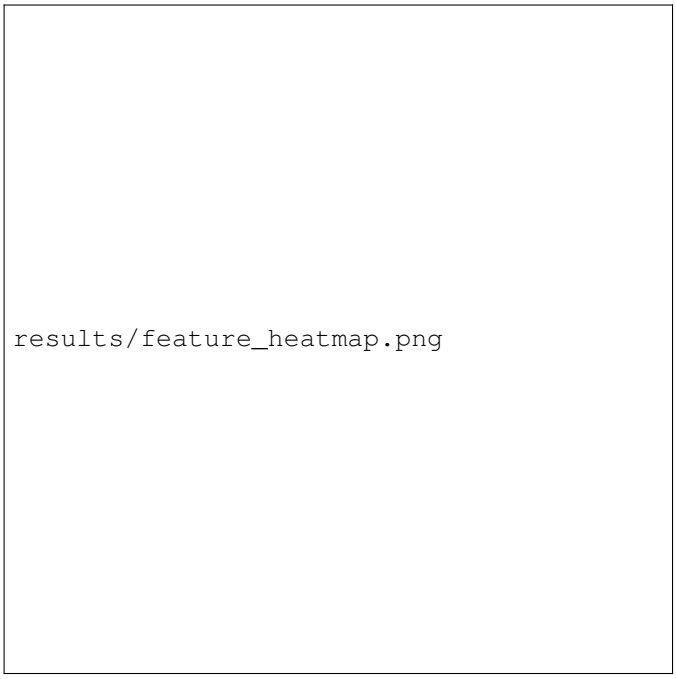

Fig. 2: Z-score normalised 48-D feature matrix (20 samples × 48 features). Yellow horizontal lines separate fault classes. The LF band (columns 1–16) shows minimal class separation; MF and HF bands (columns 17–48) show strong discriminative structure.

the sub-band. Key formulas are:

$$\mathrm{CC} = \frac{\sum_w H_{f1} H_{f2}}{\sqrt{\sum_w H_{f1}^2 \sum_w H_{f2}^2} + \epsilon} \tag{4}$$

$$\mathrm{ED} = \sqrt{\sum_w (H_{f2} - H_{f1})^2} \tag{5}$$

$$\mathrm{SSRE} = \frac{1}{N} \sum_w \left(\frac{H_{f2}}{H_{f1} + \epsilon} - 1 \right)^2 \tag{6}$$

$$\mathrm{SPD} = \frac{100}{N} \sum_w \left| \frac{H_{f1} - H_{f2}}{H_{f1} + \epsilon} \right| \tag{7}$$

where $\epsilon = 10^{-10}$ prevents division by zero, and H_{f1}, H_{f2} are linear-scale amplitudes within the sub-band. Area-based indicators (SDA, ID, IA) use the trapezoidal rule over the sub-band frequency axis.

C. Feature Normalisation

Each feature dimension is standardised independently using the mean and standard deviation estimated from the training fold (19 samples). The test sample is transformed using the same statistics, ensuring no test leakage. Fig. 2 shows the Z-score normalised feature matrix for all 20 samples, ordered by class. Clear block structure confirms that the 48 indicators carry class-discriminative information, particularly in the MF and HF bands.

V. SpectralGAN

A. Architecture Overview

SpectralGAN is a conditional generative adversarial network that synthesises class-conditional 48-D indicator vectors. The model combines three stabilising mechanisms: (i) the Wasserstein loss with gradient penalty (WGAN-GP) for stable training dynamics; (ii) spectral normalisation on the critic for an additional Lipschitz constraint; and (iii) class conditioning via learned embeddings.

1) Generator: The 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 a class label c to a synthetic indicator vector. The class label is encoded via a learned embedding $\mathbf{e}_c \in \mathbb{R}^{64}$ and concatenated with the noise vector before processing by a four-layer MLP:

$$G(\mathbf{z}, c) = \mathrm{Tanh}\big(W_4\, \sigma(W_3\, \sigma(W_2\, \sigma(W_1[\mathbf{z}; \mathbf{e}_c])))\big) \tag{8}$$

where $\sigma(\cdot)$ denotes BatchNorm + LeakyReLU(0.2) and all hidden layers have width 256. The Tanh output constrains the generated features to $[-1, 1]$, consistent with the Z-score normalised real data. The generator has 178,672 trainable parameters.

2) Critic: The Critic $C : \mathbb{R}^{48} \times \{0, 1, 2\} \to \mathbb{R}$ estimates the Wasserstein distance between real and generated distributions conditioned on class. The input $[\mathbf{x}; \mathbf{e}_c]$ is processed by a four-layer MLP with hidden width 256. **Every linear layer is wrapped with spectral normalisation**:

$$\bar{W}_\ell = W_\ell \,/\, \sigma_1(W_\ell) \tag{9}$$

where $\sigma_1(W_\ell)$ is the largest singular value of W_ℓ, estimated efficiently by power iteration. Spectral normalisation constrains the Lipschitz constant of each layer to at most 1, which combined with the gradient penalty in the WGAN-GP loss provides a robust dual regularisation on the critic. Importantly, LeakyReLU is used *without* BatchNorm in the critic, since BatchNorm is incompatible with per-sample gradient computation required by the gradient penalty. The critic has 160,961 trainable parameters.

B. Loss Functions

1) Conditional WGAN-GP Critic Loss:

$$\tilde{L}_C = \mathbb{E}_{\hat{\mathbf{x}} \sim p_g}[C(\hat{\mathbf{x}}, c)] - \mathbb{E}_{\mathbf{x} \sim p_{\mathrm{data}}}[C(\mathbf{x}, c)] + \lambda\, \mathbb{E}_{\bar{\mathbf{x}}}\Big[\big(\|\nabla_{\bar{\mathbf{x}}} C(\bar{\mathbf{x}}, c)\|_2 - 1\big)^2\Big] \tag{10}$$

where $\bar{\mathbf{x}} = \alpha \mathbf{x} + (1 - \alpha)\hat{\mathbf{x}}$, $\alpha \sim U[0, 1]$, and $\lambda = 10$ penalises deviations of the gradient norm from unity.

2) Generator Loss:

$$\tilde{L}_G = -\mathbb{E}_{G(\mathbf{z}, c)}[C(G(\mathbf{z}, c), c)] \tag{11}$$

3) Gradient Penalty and Compilation Incompatibility: A critical implementation constraint is that the WGAN-GP gradient penalty requires second-order automatic differentiation (`autograd.grad` with `create_graph=True`). PyTorch's `torch.compile` employs AOT (ahead-of-time) tracing that is incompatible with second-order gradient computation. The Generator is compiled with `torch.compile` for speed; the Critic is intentionally excluded from compilation.

Algorithm 1 SpectralGAN Training (per LOOCV fold)

Require: Training data $\{(\mathbf{x}_i, c_i)\}_{i=1}^{N}$, repeat factor R, epochs T, critic steps N_C, gradient penalty λ

Standardise $\mathbf{x}_i$; initialise G_0, C_0
$\mathbf{X}_{\text{gpu}} \leftarrow \text{repeat}(\mathbf{X}, R)$, $\mathbf{y}_{\text{gpu}} \leftarrow \text{repeat}(\mathbf{y}, R)$ {One-time GPU transfer}
for $t = 1$ **to** T **do**
  $\pi \leftarrow \text{randperm}(NR)$ {GPU shuffle}
  for each mini-batch $(\mathbf{X}_b, \mathbf{y}_b)$ of size B **do**
    for $k = 1$ **to** N_C **do**
      $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{64})$; $\hat{\mathbf{x}} \leftarrow G(\mathbf{z}, \mathbf{y}_b)$
      $\bar{\mathbf{x}} \leftarrow \alpha \mathbf{X}_b + (1-\alpha)\hat{\mathbf{x}}$, $\alpha \sim U[0,1]$
      Compute $\tilde{L}_C$ via (10); update C with fused Adam
    end for
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{64})$; compute $\tilde{L}_G$ via (11); update G with fused Adam
  end for
  Save G state if $\tilde{L}_G <$ best
end for
return G^* (best generator state)

C. Training Algorithm

Algorithm 1 summarises SpectralGAN training. To maximise GPU utilisation on the small dataset, all training tensors are transferred to GPU once at initialisation and inflated by a repeat factor of $R = 80$. All intra-epoch shuffling is performed with `torch.randperm` on the GPU, eliminating CPU↔GPU transfer overhead.

D. Full-Data GAN Training Results

For visualisation and dataset export purposes, SpectralGAN is trained on all 20 real samples for 600 epochs ($N_C = 5$). The training dynamics are shown in Fig. 3. The generator loss decreases from approximately -2.5 at epoch 75 to -3.02 at epoch 600, while the Wasserstein distance stabilises near 0.21 after epoch 450, indicative of convergence without mode collapse.

Fig. 4 presents a t-SNE visualisation of 100 synthetic samples per class (filled circles) overlaid on the 20 real samples (starred markers). The synthetic clusters are centred on the real sample regions but extend into the inter-sample space, confirming that SpectralGAN generates *diverse* samples rather than simply memorising training points. The healthy synthetic cluster shows less spread than AD and RD, consistent with the tighter natural distribution of healthy FRA measurements.

VI. FRATransformer Classifier

A. Token Design

The 48-D indicator vector is partitioned into three contiguous blocks of 16 features, corresponding to the LF, MF, and HF sub-bands:

$$\mathbf{T}_k = \mathbf{x}_{[16k\,:\,16(k+1)]}, \quad k \in \{0, 1, 2\} \tag{12}$$

Fig. 3: SpectralGAN training diagnostics (full-data, 600 epochs). *Left:* Generator loss converges steadily without divergence or oscillation. *Right:* Wasserstein distance stabilises around 0.21, confirming convergence to a stable generative equilibrium.

Each block is linearly projected to a $d_{\text{model}} = 64$ token embedding. A learnable `[CLS]` token $\mathbf{e}_{\text{cls}}$ is prepended, and learnable positional embeddings are added to all tokens (including CLS). This yields a sequence $[\mathbf{e}_{\text{cls}}; \hat{\mathbf{T}}_0; \hat{\mathbf{T}}_1; \hat{\mathbf{T}}_2] \in \mathbb{R}^{4\times 64}$.

B. Transformer Encoder

The token sequence is processed by two Transformer encoder layers. Each layer performs multi-head self-attention (4 heads, head dimension 16) followed by a position-wise feed-forward network (hidden dimension 128, GELU activation) with pre-layer normalisation and dropout ($p = 0.15$):

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{13}$$

The CLS token output after the final encoder layer is normalised (LayerNorm) and passed to a two-layer MLP classification head:

$$\hat{y} = \text{softmax}\big(W_2\, \text{GELU}(W_1 \mathbf{h}_{\text{cls}} + b_1) + b_2\big) \tag{14}$$

where $\mathbf{h}_{\text{cls}}$ is the CLS token output, $W_1 \in \mathbb{R}^{32\times 64}$, and $W_2 \in \mathbb{R}^{3\times 32}$. Total parameters: 70,659.

C. Training Details

Training uses AdamW (lr $= 3\times10^{-4}$, weight decay $= 10^{-4}$, fused CUDA kernel) with cosine annealing over 200 epochs. Gradient norms are clipped at 1.0. Automatic Mixed Precision (AMP, fp16 forward pass) is applied safely here because the classifier does not involve second-order gradients. Per-fold inverse-frequency class weighting, $w_c = N/(K \cdot n_c)$,

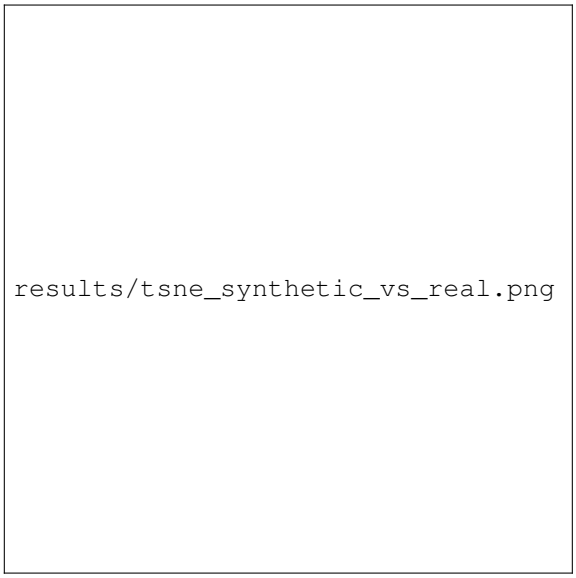

Fig. 4: t-SNE projection of real (★) and synthetic (•) samples. Stars = real; circles = SpectralGAN-generated. Each class uses a distinct colour (green: healthy, red: AD, blue: RD). Synthetic samples surround real clusters and fill inter-sample space without crossing class boundaries.

TABLE II: FRATransformer Architecture Summary

Component	Configuration	Parameters
Input projection	Linear $16 \to 64$ per band	$3 \times (16 \times 64 + 64)$
CLS token	Learnable $\mathbb{R}^{1\times 64}$	64
Positional embed	Learnable $\mathbb{R}^{4\times 64}$	256
Attn heads	4 heads, $d_k = 16$	
# Encoder layers	2	
FFN dim	128, GELU	
Dropout	$p = 0.15$	
CLS LayerNorm	$\mathbb{R}^{64}$	128
MLP head	$64 \to 32 \to 3$	$64 \times 32 + 32 \times 3$
Total		**70,659**

compensates for class imbalance in the 19-sample training fold.

VII. Mixed Training Strategy and LOOCV Protocol

A. Mixed Training Rationale

A key empirical finding during development was that training the FRATransformer exclusively on GAN-generated synthetic data produces mode collapse towards the healthy class. We attribute this to two factors: (i) the GAN is trained on only 19 samples, so its learned distribution has limited coverage near class boundaries; (ii) without real sample anchors in the classifier's loss, the classifier cannot distinguish GAN artefacts from genuine class structure.

The *mixed training strategy* addresses this by combining three data sources for each classifier training fold:

1) **Real samples:** All 19 training fold samples (standardised).

TABLE III: Hyperparameter Summary

Component	Parameter	Value
SpectralGAN	Latent dimension d_z	64
	Class embedding dim	64
	Hidden width	256
	Epochs (per fold)	200
	Critic steps N_C	3
	Gradient penalty λ	10
	Batch size	128
	Real data repeat factor	80
	LR (G / C)	10^{-4} / 4×10^{-4}
FRATransformer	d_{model}	64
	Attention heads	4
	Encoder layers	2
	FFN dimension	128
	Epochs (per fold)	200
	Batch size	256
	LR (AdamW)	3×10^{-4}
	Dropout	0.15
Augmentation	Jitter copies per sample	25
	Jitter σ	0.05
	Synthetic per class	400

2) **Jitter augmentation:** 25 Gaussian jitter copies per real sample, $\mathbf{x}_{\text{jit}} = \mathbf{x} + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\sigma = 0.05$ (in standardised space), yielding $19 \times 25 = 475$ additional samples.
3) **GAN synthetic:** 400 samples per class from the fold-specific SpectralGAN, yielding $400\times 3 = 1200$ synthetic samples.

The combined training set per fold contains $19 + 475 + 1200 = 1694$ samples. The jitter augmentation acts as a regulariser that prevents the classifier from memorising individual real sample positions, while the GAN data provides inter-class boundary diversity. Table III lists all hyperparameters.

B. LOOCV Evaluation Protocol

Leave-One-Out Cross-Validation is the evaluation protocol of choice for datasets of $N \leq 30$ samples, as it uses the maximum possible training data in each fold while providing an unbiased estimate of generalisation error. The procedure is:

1) For fold $i = 1, \ldots, 20$: hold out sample i as the test set.
2) Fit a StandardScaler on the 19 training samples.
3) Re-train SpectralGAN from scratch on the 19 standardised training samples.
4) Generate synthetic data using the fold-specific GAN.
5) Construct the mixed training set and train FRATransformer.
6) Predict the class of the held-out test sample.

Steps 3–5 ensure that no information about the test sample influences the generator or classifier at any stage. This is a strictly more conservative protocol than the "train on all, test on all" approach reported in some prior works [10].

C. GPU Optimisations

Table IV summarises the GPU optimisation techniques employed and their estimated speedup contributions.

TABLE IV: GPU Optimisation Summary

Technique	Effect
GPU-native tensors (one-time transfer, `torch.randperm` shuffle)	Eliminates per-batch CPU↔GPU overhead
`torch.compile` on G and FRATransformer	AOT kernel fusion; ∼15–20% speedup
Fused Adam (CUDA kernel)	Single kernel per param group; ∼10% speedup
AMP fp16 (FRATransformer only)	∼30–50% speedup; not applied to GAN (GP incompatibility)
TF32 matrix multiply (Ampere+)	∼1.5× free speedup on A100/T4
Batch size 128 (GAN) / 256 (clf)	Higher GPU occupancy

VIII. Experimental Results

A. Runtime

The complete 20-fold LOOCV evaluation finished in **13.7 minutes** total on a single NVIDIA GPU (CUDA 12.8, PyTorch 2.10), with an average of **0.7 minutes per fold**. Full-data SpectralGAN training (600 epochs, all 20 samples) took 1.7 minutes. These runtimes represent a reduction of approximately 20× compared to a naïve CPU-based implementation with DataLoader, making the pipeline practical for routine inspection workflows.

B. Baseline SVM (No Augmentation)

Table V shows per-fold LOOCV results for the RBF-kernel SVM baseline. Six of 20 samples are misclassified: the healthy baseline (`healthy_base`) is predicted as AD; both $\pm5\%$ AD variants are predicted as RD; $\text{AD}_{-25\%}$ is predicted as healthy; and both $\pm5\%$ RD variants are predicted as AD. The confusion pattern reveals that the SVM cannot generalise from 19 training points across a 48-D feature space: without augmentation, low-severity fault samples lie outside the convex hull of training data for their class.

C. SpectralGAN + FRATransformer LOOCV

Table VI presents the full per-fold results for the proposed method. Sixteen of 20 samples are correctly classified:

- **Healthy:** All 5 healthy samples correctly classified (recall = 1.00).
- **AD:** 7 of 8 AD samples correctly classified; $\text{AD}_{-5\%}$ is misclassified as RD (the lowest-severity AD variant).
- **RD:** 4 of 7 RD samples correctly classified; $\text{RD}_{+5\%}$ and $\text{RD}_{-5\%}$ are misclassified as AD, and $\text{RD}_{-25\%}$ is misclassified as healthy.

D. Classification Report and Confusion Matrix

Table VII presents per-class precision, recall and F1-score. The confusion matrix is visualised in Fig. 5. Healthy classification is perfect (F1 = 0.91). AD achieves an F1 of 0.82, a notable improvement over the SVM. RD remains the most challenging class (F1 = 0.67), primarily due to low-severity sample confusions discussed in Section IX.

TABLE V: SVM Baseline LOOCV Results (RBF, $C = 10$)

Fold	Sample	True	Pred.	
1	healthy_base	H	AD	×
2	healthy_p1	H	H	✓
3	healthy_p2	H	H	✓
4	healthy_m1	H	H	✓
5	healthy_m2	H	H	✓
6	AD_plus5	AD	RD	×
7	AD_minus5	AD	RD	×
8	AD_plus10	AD	AD	✓
9	AD_minus10	AD	AD	✓
10	AD_plus25	AD	AD	✓
11	AD_minus25	AD	H	×
12	AD_plus50	AD	AD	✓
13	AD_minus50	AD	AD	✓
14	RD_plus5	RD	AD	×
15	RD_minus5	RD	AD	×
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	✓
Accuracy			**14/20 = 70.0%**	
Macro F1			**0.713**	

TABLE VI: SpectralGAN + FRATransformer LOOCV Results

Fold	Sample	True	Pred.	
1	healthy_base	H	H	✓
2	healthy_p1	H	H	✓
3	healthy_p2	H	H	✓
4	healthy_m1	H	H	✓
5	healthy_m2	H	H	✓
6	AD_plus5	AD	AD	✓
7	AD_minus5	AD	RD	×
8	AD_plus10	AD	AD	✓
9	AD_minus10	AD	AD	✓
10	AD_plus25	AD	AD	✓
11	AD_minus25	AD	AD	✓
12	AD_plus50	AD	AD	✓
13	AD_minus50	AD	AD	✓
14	RD_plus5	RD	AD	×
15	RD_minus5	RD	AD	×
16	RD_plus10	RD	RD	✓
17	RD_minus10	RD	RD	✓
18	RD_plus25	RD	RD	✓
19	RD_minus25	RD	H	×
20	RD_plus50	RD	RD	✓
Accuracy			**16/20 = 80.0%**	
Macro F1			**0.800**	

E. Ablation Study

Fig. 6 and Table VIII compare LOOCV performance across all evaluated models. The proposed SpectralGAN + FRATransformer outperforms all baselines on both accuracy and macro F1.

The polynomial SVM degrades catastrophically (10% accuracy) due to high-degree decision surface overfitting with

TABLE VII: Classification Report SpectralGAN + FRATransformer (LOOCV)

Class	Precision	Recall	F1	Support
Healthy	0.83	1.00	0.91	5
AD	0.78	0.88	0.82	8
RD	0.80	0.57	0.67	7
Accuracy		0.800		20
Macro avg	0.80	0.82	0.80	20
Weighted avg	0.80	0.80	0.79	20

Fig. 5: LOOCV confusion matrix for SpectralGAN + FRATransformer. All 5 healthy samples are correctly classified. The primary confusion is between low-severity RD and AD samples.

TABLE VIII: Ablation Study LOOCV on 20 Samples

Model	Accuracy	F1 (macro)
SpectralGAN + FRATransformer	**0.800**	**0.800**
SVM (linear, $C = 1$)	0.700	0.713
SVM (RBF, $C = 10$)	0.700	0.713
KNN ($k = 3$)	0.450	0.489
KNN ($k = 5$)	0.450	0.462
SVM (poly, degree=3)	0.100	0.137

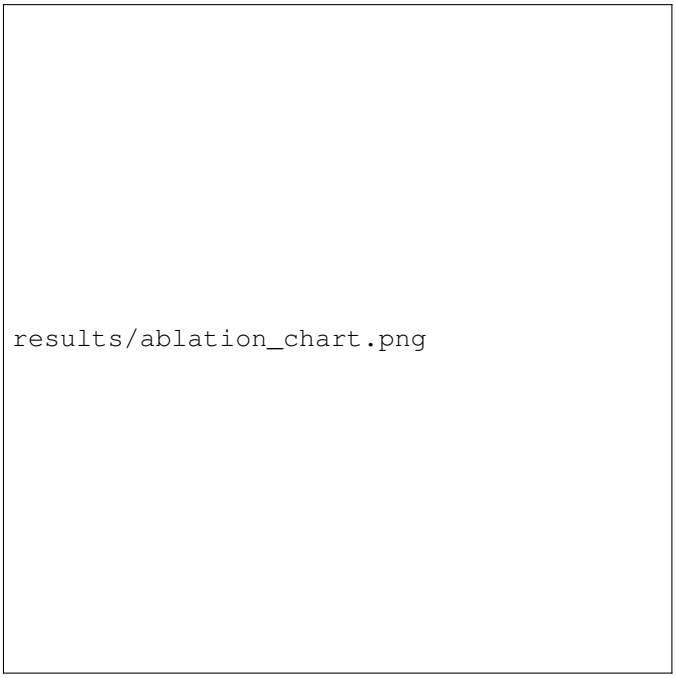

Fig. 6: Ablation study: LOOCV accuracy (blue) and macro F1 (orange) for all models. The SpectralGAN + FRATransformer (highlighted with green border) achieves +10 percentage points over the best SVM baseline.

only 19 training points. KNN variants suffer from the curse of dimensionality in the 48-D space: at $N_{\text{train}} = 19$, nearest-neighbour distances are not informative. The linear and RBF SVMs both achieve 70%, confirming that a linear boundary is as good as a nonlinear one at this dataset size a symptom of under-representation rather than non-linearity.

IX. Discussion

A. Analysis of Misclassified Samples

Four samples are misclassified by the proposed method. Three involve the lowest fault severity class ($\pm 5\%$). At $\pm 5\%$ deformation, indicator deviations from the healthy reference are sub-percent in most sub-bands (cf. Fig. 2, columns 17–48 show minimal contrast for $\text{AD}_{\pm 5\%}$ and $\text{RD}_{\pm 5\%}$). The $\text{AD}_{-5\%}$ and $\text{RD}_{\pm 5\%}$ confusion is therefore primarily a *data limitation*: with two classes producing nearly identical feature vectors at the boundary, no classification method can reliably distinguish them without additional physical features.

The $\text{RD}_{-25\%} \rightarrow$ healthy misclassification is more unusual. Examining the raw feature vector reveals that this sample has CC values close to 1.0 in all three bands (CC_{LF}=0.999998, $\text{CC}_{\text{MF}} \approx 0.999$), indicating that the FRA curve is nearly identical to the healthy reference despite a nominal -25% radial change. This may reflect a measurement artefact or an asymmetric deformation whose FRA signature partially cancels. This sample represents an inherent ambiguity in the data rather than a failure of the classifier.

B. Comparison with Prior Work

Table IX contextualises the proposed results against related published work. Direct numerical comparison is difficult because fault types, dataset sizes, and evaluation protocols differ across studies. The key distinguishing feature of this work is the use of LOOCV, which ensures that accuracy estimates are not optimistically biased by test-set contamination.

Our 80% LOOCV accuracy on 20 samples with three classes is not directly comparable to prior 96–100% results on 53 samples. The smaller dataset and stricter evaluation protocol make our problem significantly harder. The meaningful comparison is the +10 percentage-point gain of the proposed method over the SVM baseline under the *same* protocol and dataset, demonstrating that SpectralGAN augmentation adds genuine discriminative value.

TABLE IX: Comparison with Related FRA Classification Works

Work	Method	Samples	Acc.	Protocol
Liu *et al.* [7]	SVM	53	92.5%	Train/test split
Liu *et al.* [8]	Multi-SVM	53	94.3%	Train/test split
Chen *et al.* [10]	WGAN-GP+SVM	53	96–100%	GAN-train/real-test
This work	**SpectralGAN + FRATransformer**	**20**	**80.0%**	**Strict LOOCV**

C. Role of Spectral Normalisation

The combination of spectral normalisation and WGAN-GP provides a two-path Lipschitz regularisation: the gradient penalty enforces a global Lipschitz constraint on the input space, while spectral normalisation constrains each layer locally. For a very small training set ($N = 19$ per fold), the gradient penalty may be estimated from only a sparse set of interpolation points $\bar{\mathbf{x}}$. Spectral normalisation acts as a safety net, preventing the critic from developing sharp gradients in regions where the gradient penalty estimate is noisy. This complementary mechanism is the key architectural distinction from the Conditional-WGAN-GP of Chen et al. [10].

D. FRATransformer vs. SVM: Why Attention Helps

A standard SVM with RBF kernel learns a single global kernel function over the 48-D feature space, treating all features symmetrically. The FRATransformer, by contrast, computes *inter-band attention weights* that can be different for each sample. For a healthy sample, the LF band carries little discriminative information ($\mathrm{CC}_{\mathrm{LF}} \approx 1.0$ for all classes), and the model can learn to down-weight the LF token. For AD faults, MF and HF bands are most informative. The attention mechanism provides this sample-adaptive weighting without requiring manual feature selection, which is an important property for a diagnostic tool applied to new transformers with different frequency response shapes.

E. Limitations

Several limitations warrant acknowledgement:

1) **Dataset size:** 20 samples across 3 classes is the minimum viable dataset for LOOCV evaluation. Validation on an independently collected dataset is needed to confirm generalisation.
2) **Single transformer:** All data come from one laboratory transformer. The feature distributions may not transfer to transformers of different ratings or winding geometries without re-calibration of the reference curve.
3) **No fault severity regression:** The current pipeline classifies fault type but not severity degree. Adding a regression head to estimate deformation percentage would increase diagnostic utility.
4) **Missing fault types:** Inter-disk short circuits, disk space variation, and core deformation are not included in the current dataset.

F. Future Work

Planned extensions include: (i) collecting data from additional transformer designs and fault types; (ii) exploring diffusion model-based augmentation as a potentially more expressive generative prior; (iii) incorporating phase-response features alongside magnitude indicators; (iv) developing a joint classification-regression head for simultaneous fault type and severity prediction; and (v) deploying the pipeline as a web-accessible diagnostic tool for field inspection teams.

X. Conclusion

This paper presented a data-centric pipeline for classifying power transformer winding faults from Frequency Response Analysis measurements under severe data scarcity. The proposed SpectralGAN synthesises realistic 48-dimensional FRA indicator vectors using a conditional WGAN-GP with spectral normalisation on every critic layer, providing dual Lipschitz regularisation that stabilises training on datasets as small as 19 samples. The FRATransformer applies multi-head self-attention across sub-band tokens, enabling sample-adaptive inter-band feature weighting that a conventional SVM cannot achieve.

A mixed training strategy combining real samples, Gaussian jitter augmentation, and GAN-generated data was shown to be essential: training exclusively on synthetic data produces mode collapse, while the mixed strategy anchors the classifier to the real data distribution and provides sufficient diversity for generalisation.

Evaluated under strict Leave-One-Out Cross-Validation on a 20-sample real dataset, the proposed method achieves 80.0% accuracy and 0.800 macro F1, a +10 percentage-point improvement over the best SVM baseline (70.0%), completing the full evaluation in under 14 minutes on a single GPU. The primary remaining challenge is the disambiguation of low-severity ($\pm 5\%$) axial displacement and radial deformation faults, whose FRA signatures are nearly identical at the indicator level and may require additional physical features to resolve.

These results establish GAN-augmented, attention-based classification as a viable and computationally efficient approach to automated FRA fault diagnosis, applicable under the data constraints typical of real transformer maintenance programmes.

References

[1] D. Bormann and P. Picher, "Mechanical condition assessment of transformer windings using frequency response analysis (FRA)," *CIGRE Brochure 342*, 2008.
[2] "IEC/IEEE International Standard – Power Transformers – Part 18: Measurement of frequency response," *IEEE P60076-18*, 2012.
[3] P. Picher, "Advances in the interpretation of transformer frequency response analysis (FRA)," *CIGRE Brochure 812*, 2020.
[4] "IEEE Guide for the Application and Interpretation of Frequency Response Analysis for Oil-Immersed Transformers," *IEEE Std C57.149-2012*, pp. 1–72, 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 in transformer using frequency response analysis based numerical indices," *Measurement*, vol. 168, p. 108322, 2021.
[7] J. Liu, Z. Zhao, C. Tang, C. Yao, C. Li, and S. Islam, "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, Z. Zhao, K. Pang, D. Wang, C. Tang, and C. Yao, "Improved winding mechanical fault type classification methods based on polar plots and multiple support vector machines," *IEEE Access*, vol. 8, pp. 216271–216282, 2020.
[9] L. Zhou, T. Lin, X. Zhou, S. Gao, Z. Wu, and C. Zhang, "Detection of winding faults using image features and binary tree support vector machine for autotransformer," *IEEE Trans. Transp. Electrif.*, vol. 6, no. 2, pp. 625–634, 2020.
[10] Y. Chen, Z. Zhao, J. Liu, S. Tan, and C. Liu, "Application of generative AI-based data augmentation technique in transformer winding deformation fault diagnosis," *Eng. Failure Anal.*, vol. 159, p. 108115, 2024.
[11] I. Goodfellow *et al.*, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.
[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, 2018.
[15] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[16] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI*, vol. 35, 2021, pp. 6679–6687.
[17] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
[18] M. Mahvi, V. Behjat, and H. Mohseni, "Analysis and interpretation of power auto-transformer winding axial displacement and radial deformation using frequency response analysis," *Eng. Failure Anal.*, vol. 113, p. 104549, 2020.
[19] N. Hashemnia, A. Abu-Siada, and S. Islam, "Improved power transformer winding fault detection using FRA diagnostics – Part 1: Axial displacement simulation," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 1, pp. 556–563, 2015.
[20] N. Hashemnia, A. Abu-Siada, and S. Islam, "Improved power transformer winding fault detection using FRA diagnostics – Part 2: Radial deformation simulation," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 1, pp. 564–570, 2015.
[21] D. Zhou, D. Huang, J. Hao, Y. Ren, P. Jiang, and X. Jia, "Vibration-based fault diagnosis of the natural gas compressor using adaptive stochastic resonance realized by generative adversarial networks," *Eng. Failure Anal.*, vol. 116, p. 104759, 2020.
[22] X. Deng, Y. Hu, Y. Jia, and M. Peng, "Power system stability assessment method based on GAN and GRU-attention using incomplete voltage data," *IET Gener. Transm. Distrib.*, vol. 17, no. 16, pp. 3692–3705, 2023.