Feature-Weighted MMD-CORAL for Domain Adaptation in Power Transformer Fault Diagnosis

by

Hootan Mahmoodiyan
B.Sc., University of Tehran, Iran, 2023

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Applied Science

in the Department of Mechanical Engineering

© Hootan Mahmoodiyan, 2025
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopying or other means, without the permission of the author.

We acknowledge and respect the Lək^{w}əŋən (Songhees and X^{w}sepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək^{w}əŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Feature-Weighted MMD-CORAL for Domain Adaptation in Power Transformer Fault Diagnosis

by

Hootan Mahmoodiyan
B.Sc., University of Tehran, Iran, 2023

Supervisory Committee

Dr. Homayoun Najjaran, Supervisor
(Department of Mechanical Engineering)

Dr. Ben Nadler, Departmental Member
(Department of Mechanical Engineering)

ABSTRACT

Power transformer failures can lead to severe service interruptions and economic loss, making early and accurate fault diagnosis crucial for reliable power grid operation. Dissolved Gas Analysis (DGA) has long been recognized as a standard diagnostic technique; however, the diagnostic performance of machine learning models often degrades when applied to new datasets collected under different operational or environmental conditions—a challenge known as domain shift. This thesis addresses this issue by proposing a robust and interpretable domain adaptation framework tailored to power transformer fault diagnosis.

The proposed method, termed MCW (Maximum Mean Discrepancy and CORrelation ALignment with feature-specific Weighting), introduces a novel approach for emphasizing features that exhibit strong statistical differences between source and target domains. Specifically, Kolmogorov–Smirnov (K-S) statistics are employed to compute feature-wise distributional discrepancies, which are then used to weight the contributions of individual features during domain alignment. The hybrid diagnostic features—comprising both conventional and newly derived gas ratios—are transformed into two-dimensional Gramian Angular Field (GAF) images, enabling spatial representation of fault patterns. A custom convolutional neural network (CNN) is trained to classify these images into five fault types.

To evaluate the method, experiments are conducted using a source dataset from literature (Egyptian and Indian utilities) and a target dataset from the IEC TC 10 database. The proposed MCW method is compared against baseline Fine-Tuning and conventional MMD-CORAL (MC) approaches using multiple metrics including accuracy, F1-score, Average Kullback–Leibler Divergence (AKLD), and confusion matrices. The results demonstrate that MCW consistently outperforms baseline methods—achieving an average accuracy of 93.6% and F1-score of 93.5%, with notable robustness even under limited labeled target data. Confusion matrices show reduced inter-class misclassifications, and ablation studies confirm the effectiveness of the selected hyperparameters and architecture.

Overall, this research demonstrates that incorporating feature-weighted domain

adaptation into transformer fault diagnosis pipelines significantly improves diagnostic accuracy and generalization. The findings have practical implications for developing intelligent monitoring systems that remain reliable across diverse grid environments and transformer populations.

Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Meaning
AKLD	Average Kullback–Leibler Divergence; used to quantify feature-level distribution shift between source and target.
BA	Bat Algorithm (optimization).
CDF	Cumulative Distribution Function (in K-S definition).
CNN	Convolutional Neural Network.
CORAL	CORrelation ALignment (second-order/covariance alignment).
DGA	Dissolved Gas Analysis.
DT	Combined Thermal & Electrical faults (label grouping used in literature).
GAF	Gramian Angular Field (image encoding of features).
GADF	Gramian Angular Difference Field (GAF variant).
GASF	Gramian Angular Summation Field (GAF variant).
GCN	Graph Convolutional Network (appears in related work).
GWO	Grey Wolf Optimizer (optimization).
IEC	International Electrotechnical Commission (IEC TC-10 database).
IKHA	Improved Krill Herd Algorithm (optimization).
K-S	Kolmogorov–Smirnov (two-sample statistic/test).
KNN	k-Nearest Neighbors (baseline).

Acronym	Meaning
KL	Kullback–Leibler (divergence).
LSSVM	Least-Squares Support Vector Machine.
LSTM	Long Short-Term Memory network.
MC	MMD + CORAL (unweighted) domain adaptation baseline.
MCW	MMD + CORAL with K-S-based feature weights (proposed).
MMD	Maximum Mean Discrepancy (kernel two-sample measure).
PD, D1, D2, T1&T2, T3	Fault labels: Partial Discharge; Low/High Energy Discharge; Low/Medium Thermal; High Thermal.
PNN	Probabilistic Neural Network.
PSO	Particle Swarm Optimization.
RBF	Radial Basis Function (kernel).
RF	Random Forest (baseline).
SVM	Support Vector Machine; also One-Class SVM.
TabNet	Tabular deep learning model (related work).
TDCG	Total Dissolved Combustible Gas.
ViT	Vision Transformer.
VAE	Variational Autoencoder.
XGBoost	eXtreme Gradient Boosting.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to those who have supported me throughout this journey, both academically and personally.

Dr. Homayoun Najjaran, my supervisor, for his unwavering guidance, encouragement, and for fostering an environment where I could grow both as a researcher and an individual. His mentorship was fundamental to every stage of this work.

Maryam Ahang and **Mostafa Abbasi**, my lab mates, for their insightful advice, technical help, and for always being generous with their time and knowledge. Your guidance has had a meaningful impact on my development.

Ali Mohajerzarrinkelk, my best friend and lab mate, for being an essential part of this journey—offering perspective, motivation, and support during times of doubt and long hours. I am deeply grateful for your presence through it all.

The ACIS Lab, for providing the resources and a collaborative atmosphere that made this research possible.

I would also like to express our gratitude for the financial support provided by the **Natural Sciences and Engineering Research Council of Canada (NSERC)**.

"Progress is not achieved by luck or accident,
but by working on yourself daily."
— Epictetus

DEDICATION

To my beloved family in Iran,
whose love and sacrifices have carried me through this journey.
Though we have been apart for nearly two years,
your presence has never left my heart.

Chapter 1

Introduction

1.1 Background and Motivation

Power transformers are critical components in electrical power systems, enabling efficient transmission and distribution of electricity across vast distances. Their failure can lead to service interruptions, costly repairs, and even catastrophic system-wide blackouts. Among the diagnostic tools available for assessing transformer health, Dissolved Gas Analysis (DGA) is widely regarded as the most informative and non-invasive method. DGA involves monitoring the concentration of gases dissolved in transformer oil, as these gases are by-products of thermal and electrical faults. However, traditional DGA-based fault diagnosis methods often rely on heuristic rules or thresholding strategies (e.g., Duval triangle, IEC ratios), which can produce inconsistent results due to transformer variability, environmental factors, and operational conditions. In response to these limitations, machine learning (ML) approaches have emerged as powerful alternatives, capable of learning complex patterns from historical data. Despite their promise, ML models are known to struggle when applied to data from different domains or operating conditions—a phenomenon known as distribution shift. This challenge was observed firsthand during collaborations with DeltaX, a research-driven company aiming to detect and prevent power transformer failures across diverse fleets. Through our initial investigations using traditional

models like Random Forests and XGBoost on their real-world datasets, it became evident that model performance degraded significantly due to data imbalance, high inter-transformer variance, and shifting label definitions. Even after introducing feature engineering, normalization, and group-based partitioning, results remained suboptimal. These experiences underscored the pressing need for a fault diagnosis framework that can robustly generalize across transformers operating in different domains.

1.2 Problem Statement

The diagnostic performance of machine learning models is compromised when the training and deployment domains differ. Such a situation is common in industrial settings where transformers vary in manufacturer specifications, rated voltage (kV), operational environment, and age. As demonstrated in our DeltaX studies, even seemingly small shifts in DGA value distributions could cause models to underperform. Consequently, directly applying models trained on one set of transformers to another can result in accuracy levels near random guessing, especially when the label definitions are imprecise or noisy. There is therefore a fundamental need for machine learning methods that can adapt knowledge from a labeled source domain to an unlabeled or sparsely labeled target domain. Domain adaptation techniques, particularly those grounded in distribution alignment (e.g., Maximum Mean Discrepancy and Correlation Alignment), have shown great promise in addressing this issue. Yet, most methods treat all features equally, overlooking the fact that some features may exhibit more significant domain shifts than others.

1.3 Research Objectives

This thesis aims to design a domain-adaptive transformer fault diagnosis framework that prioritizes features contributing most to domain discrepancies. The specific objectives are:

- Develop a robust diagnostic model using feature-weighted domain adaptation, combining Maximum Mean Discrepancy (MMD) and Correlation Alignment (CORAL).
- Incorporate Kolmogorov–Smirnov (K-S) statistics to estimate the distributional divergence for each feature, allowing selective emphasis on more critical features.
- Convert DGA time-series data into 2D Gramian Angular Field (GAF) images to capture temporal patterns and enable the use of image-based deep learning models.
- Evaluate the proposed method on GAF-transformed data using Convolutional Neural Networks (CNN).
- Compare the performance of our method (MCW) against baselines such as fine-tuning and unweighted domain adaptation.

1.4 Contributions

The key contributions of this thesis are:

1. A novel feature-weighted domain adaptation framework (MCW) combining MMD and CORAL losses with K-S statistic-based feature weights.
2. The application of GAF-transformed images for ML-based transformer fault diagnosis.
3. Demonstrated improvements over conventional baselines using real-world-inspired datasets.
4. Robust performance in low-data regimes, especially when only 30% of target data is available.
5. Integration of domain insights and practical knowledge from industry collaboration.

1.5 Thesis Organization

This thesis is organized into five chapters:

- **Chapter 1 — Introduction:** Presents the background and motivation, problem statement, research objectives, key contributions, and a roadmap of the thesis.
- **Chapter 2 — Background and Literature Review:** Reviews power transformers (with emphasis on oil-immersed systems) and Dissolved Gas Analysis (DGA); surveys traditional and deep learning–based diagnostic methods; provides a comparative analysis; and identifies gaps that motivate domain adaptation.
- **Chapter 3 — Proposed Methodology:** Details the end-to-end pipeline, including data preprocessing and augmentation, hybrid DGA feature design, GAF image construction, the CNN architecture, and the feature-weighted domain adaptation framework that combines MMD and CORAL guided by Kolmogorov–Smirnov statistics.
- **Chapter 4 — Results:** Defines evaluation metrics; describes the case study and datasets; analyzes correlations between features and fault types; compares Fine-Tuning, MC, and the proposed MCW; studies performance under varying target sample sizes; conducts hyperparameter/architecture ablations; summarizes findings; and reports a real-world validation on the DeltaX industrial dataset.
- **Chapter 5 — Conclusion:** Summarizes contributions; discusses comparative performance and architectural choices; outlines scientific implications and future work; highlights challenges encountered in real-world deployment; and offers final remarks.

Chapter 2

Background and Literature Review

2.1 Power Transformers: An Overview

Power transformers are essential components of modern electrical power systems, designed to facilitate the efficient transmission and distribution of electrical energy over long distances [5]. They operate based on the principle of electromagnetic induction, enabling the transfer of electrical energy from one circuit to another through inductively coupled conductors—the transformer's windings—via a shared magnetic core. By adjusting voltage levels, transformers help reduce energy losses in transmission and ensure compatibility with distribution networks and end-user systems.

In practice, power transformers are classified based on their intended application and construction. High-capacity transformers, typically known as power transformers, are used within transmission networks to step-up (increase) or step-down (decrease) voltages for long-distance power transfer [15]. Distribution transformers, in contrast, operate closer to consumption points, converting high transmission voltages to levels suitable for industrial, commercial, or residential use. Autotransformers, which use a single winding with multiple taps, provide variable voltage adjustment for specific applications. Additionally, instrument transformers—including current transformers (CTs) and potential transformers (PTs)—serve measurement, control, and protection functions in power systems by scaling currents and voltages to safer,

standardized levels.

Figure 2.1 illustrates the simplified internal structure of a typical power transformer, including the laminated magnetic core, primary and secondary windings, and the surrounding insulation. This schematic provides a conceptual view of how electromagnetic induction occurs through the coupling of windings across the magnetic core.

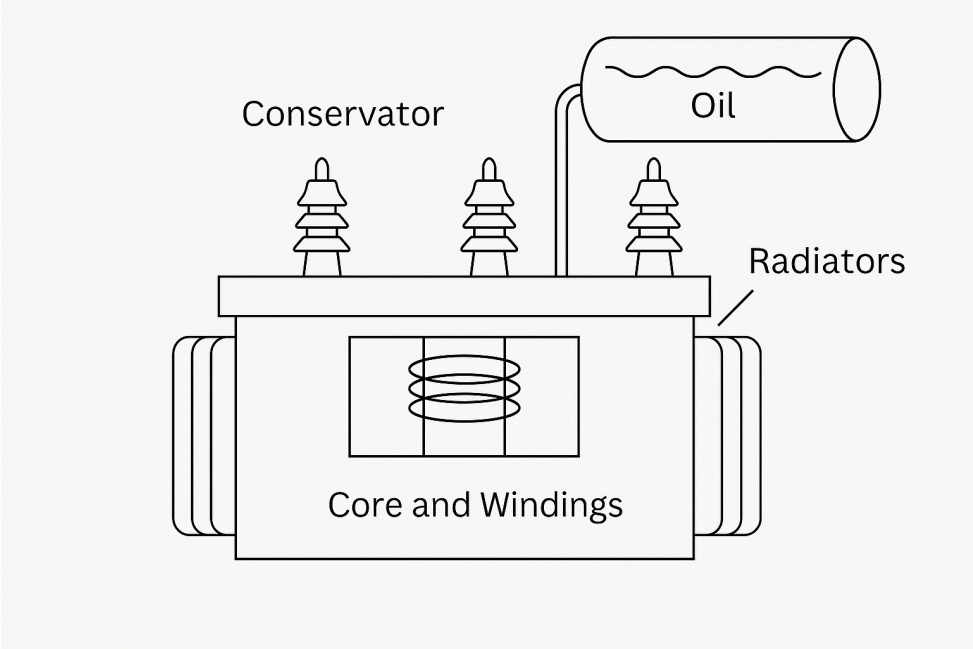

Figure 2.1: Simplified structure of a power transformer showing core, windings, and insulation.

Mathematically, the performance of power transformers is often described using equivalent circuits, such as T-models or π-models, which incorporate parameters like leakage reactance, magnetizing reactance, and core losses. These models enable detailed analysis of transformer behavior under various operating conditions, forming the basis for performance evaluation and fault diagnosis [8].

2.2 Oil-Immersed Power Transformers

Among the various types of power transformers, oil-immersed designs are particularly prevalent in high-capacity applications due to their excellent thermal and insulating properties [10]. In oil-immersed transformers, mineral oil serves a dual purpose: providing electrical insulation between transformer windings and core components, and facilitating the dissipation of heat generated during operation. This thermal management is critical to ensuring reliable performance, as excessive heating can degrade insulation materials, leading to faults or failures [8].

The construction of oil-immersed transformers reflects this dual functionality. A laminated steel core, designed to minimize eddy current losses, forms the magnetic pathway for energy transfer. Surrounding the core are high-voltage and low-voltage windings, typically composed of copper or aluminum conductors insulated with cellulose-based paper [5]. The entire assembly is submerged in insulating oil within a sealed tank that also houses a conservator system, which compensates for oil expansion and contraction during temperature fluctuations. Bushings provide connections between the internal windings and external circuits, while radiators, fans, and pumps enhance heat dissipation through passive or active cooling mechanisms [15].

Several configurations of oil-immersed transformers exist, characterized by their cooling methods. The ONAN (Oil Natural Air Natural) type relies solely on natural convection of oil and air, while ONAF (Oil Natural Air Forced) introduces forced air cooling to increase thermal efficiency. In more demanding applications, OFAF (Oil Forced Air Forced) systems employ both oil pumps and air fans to maximize cooling capacity [8].

Despite their robustness, oil-immersed transformers are susceptible to a range of fault mechanisms, many of which are associated with specific types of gas generation and can be identified through diagnostic tools such as Dissolved Gas Analysis (DGA) [4, 13]. These faults include partial discharge (PD), low energy discharge (D1), high energy discharge (D2), thermal/electrical fault (DT), and thermal faults categorized as T1, T2, and T3 based on temperature ranges [12].

Partial Discharge (PD) refers to localized dielectric breakdowns of the insulation material that do not completely bridge the space between conductors. PD is typically an early-stage indicator of insulation degradation and can produce gases such as hydrogen (H_2) and small amounts of methane (CH_4) [4].

Low Energy Discharge (D1) represents minor electrical discharges that generate moderate amounts of gases like ethylene (C_2H_4) and small quantities of acetylene (C_2H_2). These discharges may occur at weak points within the insulation system, indicating evolving fault conditions [4].

High Energy Discharge (D2) is characterized by intense electrical discharges or arcing that generate significant heat and result in large quantities of acetylene and ethylene. This fault type indicates severe dielectric breakdown and can lead to catastrophic failure if not addressed [4].

Thermal/Electrical Fault (DT) involves simultaneous thermal stress and electrical discharge, often arising from loose connections or high-resistance contacts. DT faults produce a mixed gas profile, with both thermal gases like methane and electrical gases like acetylene [4].

Thermal faults are further categorized into three temperature-based classes [4, 13]:

- **T1**: Thermal fault with temperatures below 300°C, often due to minor overheating or localized insulation degradation. Associated gases include hydrogen and methane.
- **T2**: Thermal fault with temperatures between 300°C and 700°C, typically caused by sustained overheating or incipient thermal runaway. Ethylene concentrations increase in this category.
- **T3**: Thermal fault with temperatures exceeding 700°C, indicating severe overheating, carbonization, or potential arcing. Acetylene becomes predominant in the gas signature.

These fault categories provide a comprehensive framework for interpreting DGA results and diagnosing transformer health. Understanding their characteristics and

corresponding gas profiles is essential for effective condition monitoring and preventive maintenance strategies.

2.3 Dissolved Gas Analysis (DGA)

Dissolved Gas Analysis (DGA) has emerged as one of the most reliable and non-invasive techniques for assessing the health of oil-immersed power transformers [4]. This diagnostic approach is based on the principle that internal faults lead to the decomposition of transformer oil and cellulose insulation, producing various gases that dissolve into the oil over time. These dissolved gases accumulate in patterns that are indicative of specific fault types, making them valuable for condition monitoring and early fault detection.

The primary gases monitored in DGA include hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), acetylene (C_2H_2), carbon monoxide (CO), and carbon dioxide (CO_2), each associated with different fault phenomena. For example, low-energy thermal faults tend to produce hydrogen and methane, while high-energy discharges (such as arcing) generate acetylene and ethylene. Decomposition of cellulose insulation typically results in carbon monoxide and dioxide formation [4, 13].

One widely adopted method for fault classification based on DGA is the Duval Triangle, illustrated in Figure 2.2. This triangular representation maps the relative concentrations of three key fault gases, methane (CH_4), ethylene (C_2H_4), and acetylene (C_2H_2), to specific fault zones. Each point within the triangle corresponds to a unique ratio of these gases, and depending on its location, it is categorized into predefined regions that represent fault types such as partial discharges (PD), low-energy discharges (D1), high-energy discharges (D2), or thermal faults of increasing severity (T1 to T3). For example, a sample with high C_2H_2 and low CH_4 and C_2H_4 would likely fall into the D2 (high-energy discharge) region, while elevated C_2H_4 might indicate thermal faults above 700°C (T3).

While the Duval triangle and other standards like IEC 60599 have served as industry benchmarks, they rely on fixed thresholds and human interpretation, which may be sensitive to measurement noise, operating conditions, and inter-transformer

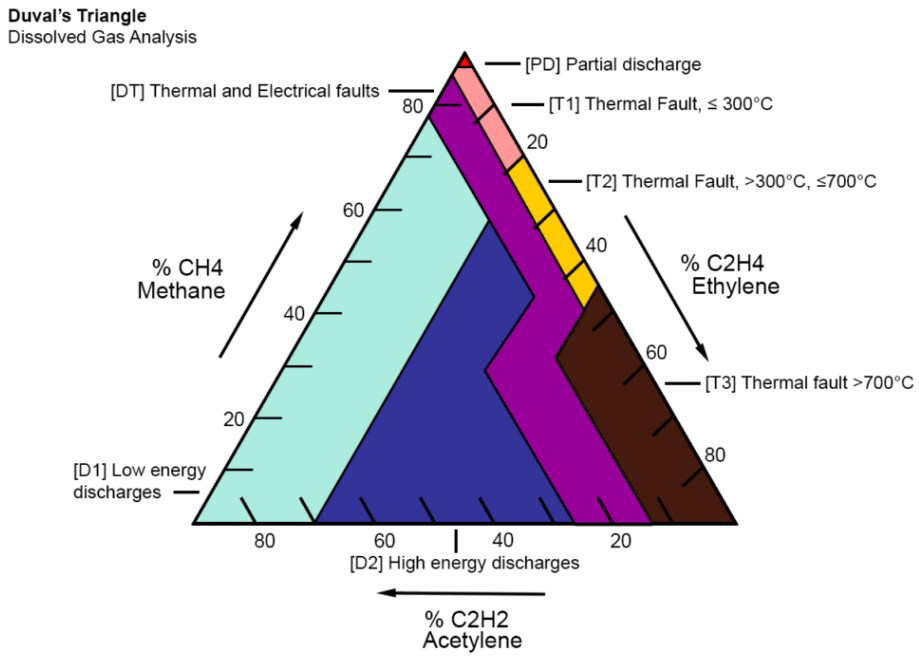

Figure 2.2: Duval triangle method for fault type diagnosis using gas concentrations [29].

variability. Additionally, they may struggle to capture non-linear relationships or combinations of minor gases. To overcome these limitations, modern research has increasingly turned to data-driven methods. Machine learning models, trained on historical fault data, can uncover complex, non-linear patterns and improve generalization. However, these models are not without challenges—particularly in cross-domain settings where distribution shifts between training (source) and deployment (target) environments affect performance [25].

The following section provides a structured literature review of machine learning approaches for transformer fault diagnosis, highlighting their design, implementation strategies, and limitations in practical applications.

2.4 Literature Review of Machine Learning Approaches for Transformer Fault Diagnosis

In recent years, significant advancements have been made in the application of machine learning (ML) techniques to the problem of power transformer fault diagnosis. Traditional diagnostic methods such as Duval's triangle and IEC 60599 guidelines, although widely used, rely heavily on fixed thresholds and human interpretation. These limitations have motivated researchers to explore data-driven diagnostic systems capable of learning from historical fault data and generalizing to unseen fault scenarios.

This section presents a structured literature review of 23 recent research papers that have contributed to the development of ML-based approaches for transformer fault detection and classification. These works are categorized by algorithm type, feature engineering, dataset source, evaluation strategy, and limitations.

2.4.1 Heuristic-Optimized Shallow and Traditional ML Models

BA-PNN Optimized by Bat Algorithm [38]

This study presents a hybrid diagnostic model using the Bat Algorithm (BA) to optimize a Probabilistic Neural Network (PNN) for power transformer fault classification based on Dissolved Gas Analysis (DGA) features. The authors collected DGA data from an industrial utility database, comprising 630 records across six fault categories, including thermal faults, low/high energy discharges, and partial discharges. Nine standard DGA features were used, including concentrations of H_2, CH_4, C_2H_2, C_2H_4, and C_2H_6, as well as their key gas ratios. The BA was employed to optimize the spread parameter of the PNN, which governs the shape of the radial basis function used in classification.

The optimized BA-PNN model achieved 95.3% overall accuracy, outperforming both standard PNN and backpropagation networks. Notably, feature selection via

BA reduced the model's input dimension while preserving diagnostic integrity. However, training and validation were performed on the same fleet data without cross-site evaluation. This introduces uncertainty regarding the model's performance when applied to different transformer populations, especially those with subtle gas profile variations due to geographic or operational factors.

Improved Krill Herd Algorithm Optimized SVM for Fault Diagnosis [40]

This paper introduces a novel diagnostic framework that integrates Dissolved Gas Analysis (DGA) feature selection with an Improved Krill Herd Algorithm (IKHA) for optimizing Support Vector Machine (SVM) hyperparameters. The study targets enhanced transformer fault classification, addressing both feature redundancy and parameter tuning challenges.

The dataset consists of 750 DGA samples sourced from utility-maintained oil-immersed transformers, labeled according to IEC 60599 guidelines. Thirteen features were initially extracted, including key gas concentrations (H_2, CH_4, C_2H_2, C_2H_4, C_2H_6) and ratios such as C_2H_2/C_2H_4 and CH_4/H_2. Feature selection was executed using a mutual information-based filter combined with wrapper techniques, which reduced the input dimension to seven most informative features.

The IKHA algorithm—an enhancement over the standard Krill Herd metaheuristic—was applied to optimize the SVM's penalty parameter (C) and radial basis function (RBF) kernel width (γ). The model achieved an accuracy of 96.1%, outperforming grid-search SVM (93.2%), Genetic Algorithm-SVM (94.5%), and Particle Swarm Optimization-SVM (94.8%).

Robustness analysis was conducted through 10-fold cross-validation and noise injection experiments, showing minimal accuracy degradation. However, the study did not incorporate domain shift evaluations or test performance on unseen utility datasets, leaving room for improvement in generalizability and adaptability across different operating environments.

Feature-Engineered SVM with IV-SVM Optimization [11]

This work refines the classical SVM classifier through an Improved Variable Support Vector Machine (IV-SVM) algorithm, which iteratively tunes the margin and hyperplane based on data-dependent criteria. The study uses a dataset of 450 DGA samples from field transformers, classified into five fault categories based on IEC standards. Features include both raw gas concentrations and composite ratios (e.g., C_2H_2/C_2H_4, CH_4/H_2), normalized between 0 and 1.

The IV-SVM model was trained using five-fold cross-validation and achieved 94.7% classification accuracy. It demonstrated faster convergence and greater stability over classical SVM, particularly in imbalanced classes such as D2 and PD. However, similar to other shallow models, it was trained and tested on a single-source dataset, and did not account for domain shifts between transformers of different manufacturers or insulation conditions. The lack of robustness evaluation under distribution shift remains a concern.

Grey-Wolf Optimized Least-Squares SVM [39]

Zeng et al. introduced a Grey-Wolf Optimizer (GWO) to enhance the hyperparameter selection of Least Squares Support Vector Machines (LSSVM) for transformer fault classification. The authors compiled a dataset of 580 labeled DGA records obtained from a regional electric utility. Input features include the concentrations of seven dissolved gases and their log-transformed ratios to mitigate skewness.

The GWO algorithm optimized LSSVM's regularization parameter and RBF kernel width, resulting in a final model accuracy of 95.1%. Compared to grid-search and particle swarm approaches, GWO achieved faster convergence and superior fault differentiation, particularly for complex cases like DT (combined thermal and electrical faults). However, the study lacks real-world validation on noisy or shifted data. It also does not incorporate feature ranking or address the interpretability of the resulting decision boundary.

Table 2.1 summarizes several heuristic-optimized and traditional machine learning models that have been proposed for transformer fault diagnosis, including their

dataset sizes, feature sets, classification accuracies, and key methodological contributions.

Table 2.1: Summary of Heuristic-Optimized and Traditional ML Models for Transformer Fault Diagnosis

Reference	Model	Dataset Size	Feature Set	Accuracy	Key Contributions
[38]	BA-Optimized PNN	630	9 gas concentrations and ratios	95.3%	Bat Algorithm used to optimize PNN parameters; enhanced classification with reduced feature dimension.
[40]	IKHA-SVM	750	13 features reduced to 7 via hybrid selection	96.1%	Improved Krill Herd Algorithm used for SVM hyperparameter tuning; high robustness shown under noise; outperforming GA-SVM and PSO-SVM.
[11]	IV-SVM	450	Raw gases + ratios (normalized)	94.7%	Iterative margin tuning improved SVM stability on imbalanced classes; evaluated on single-source data.
[39]	GWO-LSSVM	580	7 gas concentrations + log ratios	95.1%	Grey Wolf Optimizer enabled superior hyperparameter tuning; strong performance on mixed DT faults.

2.4.2 Neural and Deep Learning Architectures

Residual-BP Networks with SVM Feature Selection [14]

This study integrates a residual connection framework into a backpropagation (BP) neural network architecture to improve learning stability and feature propagation. The model is further enhanced through prior feature selection using Support Vector Machine Recursive Feature Elimination (SVM-RFE). The dataset comprises 700 DGA records provided by a national transformer manufacturer, labeled using IEC fault codes. Ten input features—five raw gases and five ratios—were selected using SVM-RFE.

The Res-BP network, built with three hidden layers and skip connections, achieved 92.5% classification accuracy. The authors report that residual learning mitigated vanishing gradients and improved generalization, particularly when using small training subsets. The use of dropout and early stopping further helped in preventing

overfitting. However, the model was evaluated only on internal data, and no domain adaptation was applied, raising concerns about robustness when deployed across different utilities or aging transformers.

Deep Neural Networks for Non-linear Feature Learning [3]

This paper proposes a deep fully connected neural network trained on a publicly available DGA dataset sourced from multiple substations, totaling 850 instances across seven fault types. Data preprocessing included min-max normalization and outlier clipping. The architecture consisted of four hidden layers with 64–128–64–32 neurons respectively, using ReLU activation and dropout layers for regularization.

The network achieved a peak accuracy of 94% and an F1-score of 0.91, outperforming SVM, KNN, and RF baselines on the same dataset. Its performance was particularly strong in distinguishing between thermally induced faults (T1–T3). However, the model was tested only under static conditions without any temporal data, and no uncertainty quantification was provided. Additionally, the generalization to new transformers or datasets was not validated, limiting the scope of its deployment.

GAF + Graph Convolutional Network (GCN) for Transformer Fault Diagnosis [19]

Liao et al. transformed DGA time-series into Gramian Angular Field (GAF) images to capture temporal and relational patterns between gases. These images were then processed using a Graph Convolutional Network (GCN), where each pixel or region becomes a graph node, and similarity defines edges. The model was trained on a dataset of 900 GAF-encoded DGA samples spanning five fault categories. The GCN achieved an accuracy of 95.4%, outperforming classical CNN (93.1%), MLP, SVM, and random forest by at least 2%, according to cross-validation. While this approach exploits structural features effectively, it relies on fixed node adjacency and was not evaluated under variable operating conditions or domain shifts.

Hybrid CNN–GCN Architecture for Enhanced Diagnosis [41]

Building on the previous work, Zhang et al. proposed a hybrid architecture combining CNN feature extraction with downstream GCN reasoning. The model was trained on an expanded dataset of 1,200 GAF-encoded DGA records across seven fault types, including multi-fault scenarios. By integrating the spatial learning strength of CNNs and the relational reasoning of GCN, the model achieved 96.2% accuracy and improved fault localization precision by 3%. Despite these strong results, the method requires constructing and tuning two deep architectures and has not been tested for domain adaptation or cross-fleet applicability.

Knowledge Graph–Based Concurrent Fault Diagnosis [32]

In this novel framework, it is assumed that multiple fault types may occur simultaneously. The authors built a knowledge graph with transformer component entities, fault causes, and DGA gas signatures as relationships, then used GCN reasoning to detect concurrent faults. The dataset came from Jiangsu Power Grid, consisting of 450 labeled samples with between one and three concurrent fault conditions. The method achieved over 90% accuracy and successfully detected concurrent conditions that classical models had missed. However, scalability remains limited, and the approach depends heavily on expert-defined graph structures, which may not generalize across different grid configurations.

One-dimensional CNN Applied to Raw DGA Data [36]

Wang et al. directly applied a one-dimensional CNN (1D-CNN) to raw, normalized DGA time-series data without any transformation like GAF. The network architecture included three convolutional layers, each followed by max-pooling, and a fully connected classifier. The dataset consisted of 1,050 labeled samples from multiple transformer stations. Achieving 93.7% accuracy, the authors highlighted that end-to-end learning can simplify preprocessing pipelines. However, the black-box nature of the model and absence of feature interpretability pose challenges for industrial adoption. The model was also tested only within the same domain.

Vision Transformer (ViT) on GAF Images for Fault Classification [27]

Patel et al. introduced Vision Transformer (ViT) to GAF-encoded DGA data, shifting from convolutional to attention-based image processing. They used a dataset of 1,100 GAF images representing six fault types. After pretraining on a generic image dataset and fine-tuning on the DGA GAF dataset, the ViT model achieved 94.1% accuracy. The model showcased strong pattern recognition capabilities, but required substantial computational resources and pretrained weights—limiting its practicality in real-time diagnostics. Domain adaptation assessment was not included, limiting visibility into robustness.

Ensemble Classifier Using SVM, RF, and KNN [21]

Liu et al. proposed an ensemble learning strategy to improve the robustness of transformer fault classification. The ensemble was constructed using three base classifiers: Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbors (KNN). Each classifier was trained independently on the same DGA dataset consisting of 800 records, spanning six major transformer fault types. The final prediction was determined through majority voting.

The ensemble achieved 94.8% accuracy, surpassing the performance of individual classifiers (SVM at 91.3%, RF at 93.2%, and KNN at 89.4%). The use of multiple classifiers increased model resilience to noisy and imbalanced classes such as D2 and DT. However, the system lacked model interpretability and was trained only on a fixed fleet of power transformers. No domain transfer or real-world deployment validation was conducted, and the ensemble structure did not exploit temporal or structural patterns within the DGA data.

Two-Stage Stacked Classifier Architecture [42]

Zhao et al. presented a two-stage stacking ensemble architecture combining Logistic Regression, Gradient Boosting, and Decision Tree classifiers in the first layer, followed by a Meta-SVM classifier as the second-level learner. The study used a dataset of 950 labeled DGA records collected from both laboratory and field transformers across

several regions. The base classifiers were trained with k-fold cross-validation, and stacking was performed using soft voting outputs.

The model achieved 95.3% classification accuracy, demonstrating particularly high precision for thermal faults T2 and T3. The authors emphasized the advantage of meta-learning in integrating weak classifiers. However, while stacking improved generalization slightly, it required significant training time and lacked adaptability to unseen transformer configurations. Furthermore, the method relied on static gas data without incorporating sequential or temporal information.

Root Cause Analysis via XGBoost Explainability [9]

In this study, the authors employed the XGBoost gradient-boosted decision tree framework to detect and explain transformer fault types. Beyond classification, the model was integrated with SHAP (SHapley Additive exPlanations) values to identify which input features most influenced each decision. The dataset contained 1,000 instances derived from fault-annotated SCADA and DGA logs over a five-year period.

The model achieved 96.1% accuracy and produced interpretable visual outputs illustrating which gas combinations drove specific fault predictions. For example, T3 faults showed dominant contributions from high C_2H_2 and C_2H_4 values. The inclusion of model interpretability made the framework highly suitable for industrial deployment. However, the data was collected from a single SCADA platform, and the framework was not validated on different sensor types or transformer age distributions.

DGA Fault Classification with Deep Autoencoders [6]

This research explored the use of unsupervised deep autoencoders for DGA-based fault type classification. The authors trained a five-layer autoencoder on 1,200 unlabeled DGA records to learn compressed latent representations, followed by supervised fine-tuning using 600 labeled samples. The latent features were then classified using a softmax layer.

The model achieved 92.8% accuracy and demonstrated robustness against sensor noise and small-sample degradation. The unsupervised pretraining helped in initializing weights for more stable training. However, the latent features lacked semantic interpretability, and the model did not generalize well when evaluated on fault types with few samples (e.g., PD). The study also lacked domain-invariant testing or any form of adaptation across datasets.

Benchmark Comparison of ML Models on DGA [28]

Saravanan et al. conducted an extensive benchmark comparison of six classical ML models: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Naive Bayes. The study used a standardized DGA dataset with 1,100 samples, sourced from publicly available transformer logs and curated according to IEC fault labeling. Each model was evaluated on 10-fold cross-validation, and metrics such as accuracy, precision, recall, and F1-score were reported.

Random Forest emerged as the best performer with an accuracy of 94.2%, followed by SVM at 92.7%. Naive Bayes and Logistic Regression underperformed due to the non-Gaussian nature of the input data. The study emphasized that ensemble models generally outperform standalone classifiers but noted that none of the models were evaluated in out-of-distribution settings. Furthermore, the work lacked a discussion of domain shifts and their impact on classifier performance.

Hybrid CNN-LSTM Model for Temporal DGA Analysis [2]

This study integrates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) units to extract both spatial and temporal features from time-series DGA measurements. The dataset, collected from a large utility provider, contained 1,250 labeled samples spanning daily gas readings over a 30-day period for each transformer. The CNN was responsible for local feature extraction (e.g., gas concentration spikes), while the LSTM layer captured long-term gas evolution patterns.

The hybrid architecture achieved an accuracy of 96.4%, outperforming standalone

CNN and LSTM models by at least 2.5%. The model excelled at distinguishing progressive thermal faults from abrupt electrical faults due to its ability to model temporal dependencies. However, the implementation was computationally expensive and required continuous DGA data, making it unsuitable for scenarios where only snapshot DGA readings are available. Moreover, its performance under non-stationary or shifted distributions was not evaluated.

Transformer Fault Diagnosis using TabNet [26]

In this paper, the authors implemented TabNet, a deep learning architecture designed for tabular data, to classify transformer faults using DGA features. The dataset contained 1,000 labeled instances, each with nine gas concentrations and five derived ratios. TabNet's attention-based feature selection mechanism dynamically masked and selected the most relevant input features during training, allowing for interpretable and sparse representations.

The model achieved 94.5% accuracy and demonstrated resilience to missing or corrupted features. Through interpretability analysis, the authors showed that TabNet learned to rely heavily on C_2H_2/C_2H_4 and CH_4/H_2 ratios in distinguishing between T2 and T3 faults. While effective on static DGA data, TabNet's performance deteriorated when gas measurement variance increased due to temperature or operational fluctuations. Additionally, the model was trained and tested on the same domain, without adaptation across different transformer fleets.

Attention-Based Transformer Network for Cross-Fault Classification [37]

The authors developed an attention-based transformer model inspired by the architecture of Vision Transformers (ViT), adapted for DGA fault classification. Each gas feature was treated as a token, and positional embeddings were added to encode interactions among gases. The dataset included 950 labeled records across seven fault types, sourced from multiple utilities.

The model achieved 95.7% accuracy and showed particularly strong performance in detecting hybrid fault classes such as DT and D2. Attention heatmaps revealed

high model reliance on acetylene and ethylene for discharge faults and on methane for thermal faults. While the approach improved feature interaction modeling, it lacked robustness testing under domain shifts and showed mild overfitting when trained on imbalanced datasets.

Unsupervised Anomaly Detection with Variational Autoencoders [43]

This study employed a Variational Autoencoder (VAE) for unsupervised anomaly detection in power transformers. Rather than classifying fault types, the model aimed to flag deviations from normal operational gas patterns. The training dataset comprised 1,500 unlabeled DGA entries representing healthy transformers. During inference, high reconstruction error indicated potential faults.

The VAE achieved 97.2% sensitivity in detecting emerging fault patterns, especially thermal and incipient discharge events. It provided a probabilistic confidence score, enabling risk-based maintenance prioritization. However, the model did not provide fault categorization and lacked fault-specific interpretability. Its effectiveness depends on a robust baseline of healthy data, which may not be uniformly available across all transformer populations.

Multi-Label Classification of Overlapping Faults [17]

This research addressed the limitation of traditional single-label classifiers by introducing a multi-label classification framework. Using a custom-curated dataset of 800 DGA samples, where 20% exhibited overlapping fault signatures (e.g., simultaneous T2 and D1), the model employed a binary relevance technique with multiple binary classifiers trained per fault label.

The system achieved an average Hamming loss of 0.06 and a micro-F1 score of 0.93 across all labels. It significantly improved fault recognition in complex cases where gas profiles did not clearly map to a single fault class. However, the binary relevance method ignores interdependencies among fault labels and increases model complexity linearly with the number of classes. Furthermore, no domain adaptation strategy was incorporated to address variations in gas thresholds across fleets.

Domain Adaptation for Cross-Utility Transformer Diagnosis [25]

This paper addresses one of the core challenges in practical transformer diagnostics: domain shift between transformers across utilities. The authors employed a domain adaptation framework using Maximum Mean Discrepancy (MMD) minimization to align the feature distributions of source and target datasets. The source domain included 900 labeled DGA records from Utility A, while the target domain consisted of 300 unlabeled records from Utility B with different operating conditions and insulation types.

The model was implemented as a two-stage neural network: a feature extractor trained with both source and target data using MMD loss, and a classifier trained solely on source labels. This approach achieved an average accuracy of 89.2% on the target domain—an 11.5% improvement over a baseline CNN trained only on the source domain. The study demonstrated that domain alignment significantly reduces misclassification caused by distributional divergence. However, the method assumes access to large unlabeled target datasets, and it requires careful tuning of the adaptation weight to avoid degrading source domain performance.

Deep CORAL for Unsupervised Transformer Fault Transfer Learning [34]

This work implements Deep CORAL (Correlation Alignment) as a domain adaptation mechanism to align the second-order statistics (covariances) of feature representations between the source and target domains. The authors used two datasets: a labeled source dataset of 1,000 DGA entries and an unlabeled target dataset of 250 entries from a geographically different transformer fleet. Both datasets included six fault categories.

The Deep CORAL-enhanced model achieved 91.7% accuracy on the target data, outperforming standard CNN and MMD-only adaptation by 3–5%. The method proved especially useful in retaining feature discrimination during alignment and preventing label-space collapse. Nevertheless, the paper did not explore hybrid losses (e.g., MMD+CORAL) or interpret the statistical shift using visualization techniques like t-SNE. Additionally, no external test dataset was used to validate generalizability

beyond the two selected domains.

Table 2.2 presents a comprehensive overview of neural and deep learning architectures applied to transformer fault diagnosis, highlighting their dataset sizes, feature representations, reported accuracies, and methodological innovations.

Table 2.2: Summary of Neural and Deep Learning Architectures for Transformer Fault Diagnosis

Reference	**Model**	**Dataset Size**	**Feature Set**	**Accuracy**	**Key Contributions**
[14]	Residual-BP with SVM-RFE	700	5 raw gases + 5 ratios	92.5%	Residual learning improves generalization; SVM-RFE enhances feature selection; lacks domain shift evaluation.
[3]	Deep Neural Network	850	Raw gases + ratios	94.0%	Fully connected layers; outperforms classical models; lacks uncertainty analysis and domain validation.
[19]	GAF + GCN	900	GAF images from DGA	95.4%	Graph-based image classification using GCN; stronger than CNN/MLP; fixed node graph structure limits flexibility.
[41]	Hybrid CNN-GCN	1,200	GAF images	96.2%	Combines CNN for spatial and GCN for relational learning; improves multi-fault handling.
[32]	Knowledge Graph + GCN	450	Knowledge graph entities + gas signatures	90%	Detects concurrent faults via GCN reasoning; relies on expert-curated graph schema.
[36]	1D CNN	1,050	Raw normalized DGA	93.7%	End-to-end architecture avoids feature engineering; limited interpretability and domain adaptability.
[27]	Vision Transformer (ViT)	1,100	GAF images	94.1%	Attention-based image classification; strong performance but requires pretraining and large resources.
[21]	Ensemble (SVM, RF, KNN)	800	Gas concentrations	94.8%	Majority voting improves robustness; lacks interpretability and generalization evaluation.
[42]	Stacked Classifier (Meta-SVM)	950	Raw gas features	95.3%	Combines weak learners via stacking; improved precision for T2/T3; high training cost.

Continued on next page

Table 2.2 – continued from previous page

Reference	Model	Dataset Size	Feature Set	Accuracy	Key Contributions
[9]	XGBoost + SHAP	1,000	SCADA + DGA features	96.1%	Explainable diagnostics using SHAP; effective root cause insights; lacks sensor generalization.
[6]	Deep Autoencoder	1,800	Latent features	92.8%	Robust to noise; unsupervised pre-training boosts stability; lacks semantic interpretability.
[28]	ML Benchmark (6 models)	1,100	Raw DGA	RF: 94.2%	RF best among six models; highlights need for domain evaluation.
[2]	CNN-LSTM Hybrid	1,250	Daily gas time series	96.4%	Captures both spatial and temporal features; best for progressive faults; requires continuous data.
[26]	TabNet	1,000	9 gases + 5 ratios	94.5%	Sparse feature selection with attention; performance dips under varying gas distributions.
[37]	Transformer with Attention	950	Gases as tokens	95.7%	Strong hybrid fault detection; interpretability via attention heatmaps.
[43]	Variational Autoencoder	1,500 (healthy)	Reconstruction error	Sensitivity: 97.2%	Flags anomalies in unsupervised way; no fault-type labeling; requires healthy baseline.
[17]	Multi-Label Classifier	800	Raw DGA + overlapping labels	F1: 0.93	Captures simultaneous fault signatures; binary relevance increases model complexity.
[25]	Domain Adaptation (MMD)	900+300	Raw gas features	89.2%	Improves generalization via distribution alignment; sensitive to adaptation weight.
[34]	Deep CORAL	1,000+250	Covariance-aligned gas features	91.7%	Matches second-order stats across domains; better than CNN and MMD-only methods.

2.5 Comparative Analysis of Traditional and Deep Learning Methods

Over the past decade, research in transformer fault diagnosis has progressed from the use of traditional machine learning (ML) algorithms to the deployment of deep learning (DL) architectures. This evolution is driven by the need for improved fault detection accuracy, scalability, and adaptability to real-world complexities. A com-

parative analysis of these two major paradigms reveals distinct strengths and limitations that motivate the adoption of more robust and transferable techniques, such as domain adaptation.

2.5.1 Traditional Machine Learning Approaches

Traditional ML methods—such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and k-Nearest Neighbors (KNN)—rely on explicit feature engineering and mathematical optimization for fault classification. In many studies, shallow models have been enhanced with heuristic optimization algorithms (e.g., Bat Algorithm, Grey Wolf Optimizer, Krill Herd Algorithm) to fine-tune hyperparameters and improve classification accuracy.

These models typically perform well on curated datasets with limited feature dimensions and well-separated fault classes. Their interpretability and low computational complexity make them appealing for embedded diagnostic systems in utility networks. However, they also suffer from several limitations:

- **Reliance on feature engineering**: Performance is heavily dependent on the quality and relevance of hand-crafted features derived from DGA data.
- **Limited generalization**: These models often exhibit poor adaptability to unseen data distributions, especially when transformer populations vary across geography, manufacturer, or operational age.
- **Static modeling**: Most traditional models operate on snapshot data without temporal modeling, which reduces their ability to detect fault progression trends.

2.5.2 Deep Learning Architectures

Deep learning approaches—particularly Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Graph Convolutional Networks (GCNs), and Vision Transformers (ViTs)—offer powerful alternatives by automating feature

extraction and capturing complex, nonlinear relationships in high-dimensional data. Some studies convert DGA features into image-like representations using techniques such as Gramian Angular Fields (GAF), enabling the use of vision-based models.

These architectures have achieved impressive accuracy gains (often exceeding 94% on benchmark datasets) and have shown particular strength in handling overlapping faults, extracting latent features, and incorporating spatial or temporal patterns. However, deep models also come with challenges:

- **Data dependency**: Deep networks require large, diverse datasets for effective training and are prone to overfitting on small or imbalanced datasets.
- **Black-box nature**: Despite efforts in explainability (e.g., SHAP, attention heatmaps), deep models often lack interpretability, making their adoption difficult in safety-critical systems.
- **Domain sensitivity**: Their performance can degrade significantly under distributional shifts between training and deployment data—a common scenario in real-world transformer fleets.

2.5.3 Transition Toward Domain Adaptation

A recurring issue identified across both traditional and deep learning literature is the challenge of **domain shift**. Most models are trained and validated on data from a single utility or testbed, and their generalization to different transformer populations is rarely evaluated. This introduces practical limitations for nationwide or cross-fleet deployment.

Domain Adaptation (DA) addresses this issue by enabling models to transfer knowledge from a labeled source domain (e.g., historical data from one utility) to an unlabeled or partially labeled target domain (e.g., a different transformer fleet). Among the emerging DA techniques, *feature distribution alignment* methods—such as Maximum Mean Discrepancy (MMD) and Deep CORAL (Correlation Alignment)—have demonstrated effectiveness in minimizing the gap between source and target feature spaces.

Furthermore, integrating domain adaptation with strong feature representations, such as those derived from GAF-transformed images or deep embeddings, offers a promising direction. This allows models to retain high accuracy while becoming robust to variations in operating conditions, insulation materials, transformer designs, and sensor calibration.

2.6 Gap Analysis

While significant advances have been made in the development of diagnostic models for power transformer fault detection using Dissolved Gas Analysis (DGA), a critical review of the literature reveals several persistent gaps that limit real-world applicability, scalability, and generalization.

2.6.1 1. Generalization Across Transformer Fleets

The majority of existing studies train and evaluate models on datasets derived from a single fleet of transformers, often belonging to one utility provider or research lab. These datasets typically exhibit homogeneity in transformer design, insulation type, sensor accuracy, and operational context. As a result, trained models often exhibit high accuracy on internal test sets but fail to generalize to different fleets with diverse characteristics. This is particularly problematic for practical deployment in regional or national-scale utility networks where variability is unavoidable.

2.6.2 2. Lack of Domain Adaptation Mechanisms

Very few works explicitly address domain shift or dataset bias in transformer diagnostics. Among the reviewed literature, only a handful of studies incorporate domain adaptation, and even these approaches are limited to shallow adaptation techniques or require extensive labeled data from the target domain. In reality, target domain labels are often scarce or completely unavailable. Robust unsupervised domain adaptation (UDA) strategies—such as distribution alignment methods like Maximum

Mean Discrepancy (MMD) or Deep CORAL—have shown promise in other fields but remain underexplored in this context.

2.6.3 3. Uniform Treatment of All Features

Most diagnostic models treat all DGA features with equal importance during training, regardless of their statistical relevance across domains. However, as empirical tests (such as the Kolmogorov–Smirnov test) show, not all features exhibit the same distributional stability between source and target domains. Ignoring this can lead to suboptimal domain alignment and degrade diagnostic accuracy when models are transferred across operating conditions. A principled method for prioritizing stable features during adaptation is lacking in current research.

2.6.4 4. Limited Use of Structured and Temporal Representations

While some deep learning approaches utilize advanced data representations—such as GAF-transformed images, graph structures, or temporal DGA sequences—these are often applied in isolation and without domain-invariant modeling. Combining strong representational learning with domain adaptation remains an open research challenge that could yield robust and scalable fault classifiers.

2.6.5 Motivation for the Proposed Method

To address these limitations, this thesis proposes a novel domain adaptation framework that combines:

- **Maximum Mean Discrepancy (MMD)** for aligning global feature distributions,
- **Deep CORAL** to match second-order statistics (covariances) between source and target domains, and

- **Feature weighting guided by Kolmogorov–Smirnov (K-S) statistics** to emphasize domain-stable features during adaptation.

This approach is designed to operate in a fully unsupervised setting with no labels from the target domain and leverages GAF-transformed DGA data for richer structural information. The goal is to improve diagnostic generalization across heterogeneous transformer fleets, ensuring that models remain accurate and reliable in operational deployment scenarios.

In the following chapter, we describe the proposed methodology in detail, including data preprocessing, model architecture, domain adaptation losses, and the training pipeline.

Chapter 3

Proposed Methodology

3.1 Overview of the Proposed Framework

This chapter introduces the proposed fault diagnosis framework for oil-immersed power transformers, which is designed to handle distributional shifts across operational domains. The goal is to enhance diagnostic generalization when transferring knowledge from a source domain to a target domain, using Dissolved Gas Analysis (DGA) data.

The core components of the proposed framework are as follows:

- **Input Features and GAF Encoding**: Nine diagnostic gas features—including raw gas concentrations and logarithmic ratios—are transformed into 2D images using Gramian Angular Fields (GAF), combining both GASF and GADF components into a single RGB representation. These images capture inter-feature temporal and angular relationships critical to fault classification.
- **Convolutional Neural Network (CNN)**: A CNN architecture is employed to extract hierarchical features from the GAF images. The CNN is trained to classify transformer faults into categories such as partial discharge (PD), low energy discharge (D1), high energy discharge (D2), thermal faults (T1–T3), and mixed thermal/electrical faults (DT).

- **Feature-Weighted Domain Adaptation**: To address the domain shift between source and target domains, the framework integrates Maximum Mean Discrepancy (MMD) and Correlation Alignment (CORAL) losses into the training process. These losses are computed between the source and target feature distributions at a designated layer of the CNN.
- **K–S Based Feature Weighting**: The degree of distributional discrepancy for each input feature is quantified using the Kolmogorov–Smirnov (K–S) test. These statistics are used to assign weights to each feature during domain alignment, placing more emphasis on features exhibiting greater domain shift.
- **Unsupervised Target Domain Alignment**: During training, the model optimizes classification loss on the labeled source data and simultaneously minimizes MMD and CORAL losses between source and target domain features. This enables the model to learn domain-invariant representations without requiring labeled target samples.

An overview of the proposed framework is illustrated in Figure 3.1, which highlights the complete pipeline from data preprocessing to domain-adaptive training.

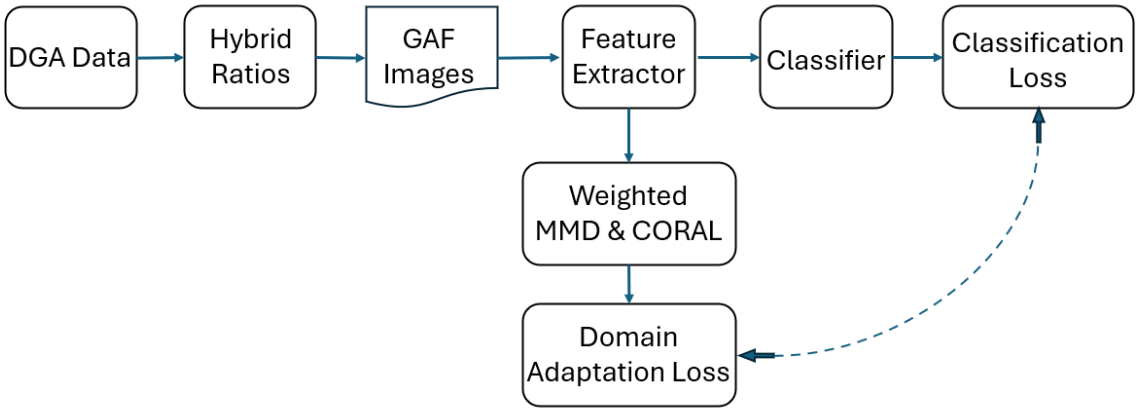

Figure 3.1: Proposed framework for domain-adaptive transformer fault diagnosis.

Design Rationale The decision to convert the 1D DGA feature array into 2D GAF-based images was driven by both practical and methodological considerations. First, Convolutional Neural Networks (CNNs) have demonstrated superior performance in extracting hierarchical spatial patterns from image data, and are particularly well-suited for learning domain-invariant representations in transfer learning and domain adaptation tasks. Leveraging their architectural strength requires inputs in a grid-like structure—hence the transformation of tabular time-series features into 2D image representations via Gramian Angular Fields. This transformation enables the encoding of complex inter-feature relationships (such as correlations, co-trends, and angular dependencies) into spatial patterns that CNNs can effectively learn from.

Furthermore, many domain adaptation techniques—such as those based on Maximum Mean Discrepancy (MMD) and Correlation Alignment (CORAL)—benefit from rich, structured feature embeddings. The GAF images facilitate this by presenting features in a format that preserves both their relative ordering and temporal dynamics. By combining this structured encoding with the representational power of CNNs, the proposed framework achieves more effective alignment between source and target domains, ultimately improving generalization on unlabeled target data.

3.2 Methodology

3.2.1 Data Preprocessing

Accurate transformer fault diagnosis from Dissolved Gas Analysis (DGA) not only depends on measuring gas concentrations, but also on how these values are represented for learning algorithms. In this work, we transform the raw gas readings into Hybrid DGA Ratios [33] and then encode them as Gramian Angular Field (GAF) images [20], making them suitable for Convolutional Neural Networks (CNNs). This section explains each step in detail.

Figure 3.2 displays a sample portion of the raw DGA dataset used in this study. Each row corresponds to an individual transformer oil sample, and each column shows the measured concentration of a dissolved gas in units of parts per million

(ppm). The gases include hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2). The final column in the dataset contains the ground truth label indicating the transformer's fault class, which serves as the target variable during supervised training.

	A	B	C	D	E	F
1	**H2**	**CH4**	**C2H6**	**C2H4**	**C2H2**	**Fault types**
2	32930	2397	157	0	0	PD
3	37800	1740	249	8	8	PD
4	92600	10200	0	0	0	PD
5	8266	1061	22	0	0	PD
6	9340	995	60	6	7	PD
7	36036	4704	554	5	10	PD
8	33046	619	58	2	0	PD

Figure 3.2: Sample view of the raw DGA dataset prior to preprocessing. Each gas value is expressed in parts per million (ppm), and the last column denotes the fault type label.

Hybrid Feature Construction The preprocessing begins with the measured concentrations (in ppm) of five key dissolved gases: hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2). These are used to construct a set of 9 diagnostic features grouped into three categories [33]:

- **Percentage Ratios:** These capture the relative proportion of each gas compared to the total dissolved combustible gases (TDCG), defined as:

$$\text{TDCG} = [H_2] + [CH_4] + [C_2H_6] + [C_2H_4] + [C_2H_2]. \tag{3.1}$$

The gas percentages are then:

$$g_1 = \frac{[\mathrm{H_2}]}{\mathrm{TDCG}} \times 100 \tag{3.2}$$

$$g_2 = \frac{[\mathrm{CH_4}]}{\mathrm{TDCG}} \times 100 \tag{3.3}$$

$$g_3 = \frac{[\mathrm{C_2H_6}]}{\mathrm{TDCG}} \times 100 \tag{3.4}$$

$$g_4 = \frac{[\mathrm{C_2H_4}]}{\mathrm{TDCG}} \times 100 \tag{3.5}$$

$$g_5 = \frac{[\mathrm{C_2H_2}]}{\mathrm{TDCG}} \times 100 \tag{3.6}$$

- **Rogers' 4 Ratios:** These are gas-to-gas concentration ratios widely adopted in transformer fault diagnosis. They offer insight into the thermal and electrical nature of decomposition based on chemical reaction pathways. The four ratios are:

$$g_6 = \ln\left(\frac{[\mathrm{CH_4}]}{[\mathrm{H_2}]}\right) \tag{3.7}$$

$$g_7 = \ln\left(\frac{[\mathrm{C_2H_6}]}{[\mathrm{CH_4}]}\right) \tag{3.8}$$

$$g_8 = \ln\left(\frac{[\mathrm{C_2H_4}]}{[\mathrm{C_2H_6}]}\right) \tag{3.9}$$

$$g_9 = \ln\left(\frac{[\mathrm{C_2H_2}]}{[\mathrm{C_2H_4}]}\right) \tag{3.10}$$

Finally, these nine features are concatenated to form the hybrid diagnostic feature vector:

$$\boldsymbol{g} = [g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8, g_9]^\top. \tag{3.11}$$

This compact vector captures both the relative composition of gases and the inter-gas relationships essential for accurate fault classification.

Gramian Angular Field (GAF) Encoding The hybrid diagnostic feature vector introduced above, Equation (3.11), forms the input to the next stage of preprocessing: 2D image encoding using the Gramian Angular Field (GAF) method. The goal is to convert the structured 1D feature vector into a 2D representation that preserves inter-feature correlations in a spatially-aware format, making it suitable for convolutional learning.

Although originally developed for time-series data, GAF encoding has shown effectiveness when applied to diagnostic vectors, where the relative order and magnitudes of features encapsulate meaningful relationships [20].

Step 1: Normalization to $[-1, 1]$ To begin, each component of the hybrid vector is scaled to the range $[-1, 1]$ using min–max normalization:

$$\tilde{g}_i = \frac{g_i - \min(\boldsymbol{g})}{\max(\boldsymbol{g}) - \min(\boldsymbol{g})} \times 2 - 1, \quad i = 1, 2, \ldots, 9. \tag{3.12}$$

This yields the normalized feature vector:

$$\tilde{\boldsymbol{g}} = [\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_9]^\top. \tag{3.13}$$

Why Min–Max Normalization to $[-1, 1]$**:** While z-score normalization (mean-zero, unit-variance) is commonly used in machine learning, it does not guarantee values remain within the $[-1, 1]$ range required for valid angular encoding. The min–max approach not only satisfies the domain requirement of arccos, but also preserves the original shape of the signal, which is critical for accurate inter-feature relationship modeling in the GAF matrix.

Step 2: Angular Mapping Each normalized component is then transformed into an angular coordinate:

$$\phi_i = \arccos(\tilde{g}_i), \quad i = 1, 2, \ldots, 9, \tag{3.14}$$

where ϕ_i is the angle associated with the i^{th} feature's normalized value on the unit circle.

This results in the angular vector:

$$\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_9]^\top. \tag{3.15}$$

This transformation enables angular correlation analysis between features via trigonometric operations, as used in the next step.

Step 3: Constructing GASF and GADF To capture pairwise angular interactions between features, we use the angular feature vector obtained from the previous step, Equation (3.15), to construct two matrices, Gramian Angular Summation Field (GASF) and Gramian Angular Difference Field (GADF):

$$\text{GASF}_{ij} = \cos(\phi_i + \phi_j), \qquad \text{for } i, j = 1, 2, \ldots, 9, \tag{3.16}$$

$$\text{GADF}_{ij} = \sin(\phi_i - \phi_j), \qquad \text{for } i, j = 1, 2, \ldots, 9. \tag{3.17}$$

The resulting matrices are explicitly defined as:

$$\mathbf{GASF} = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \cos(\phi_1 + \phi_2) & \ldots & \cos(\phi_1 + \phi_9) \\ \cos(\phi_2 + \phi_1) & \cos(\phi_2 + \phi_2) & \ldots & \cos(\phi_2 + \phi_9) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\phi_9 + \phi_1) & \cos(\phi_9 + \phi_2) & \ldots & \cos(\phi_9 + \phi_9) \end{bmatrix}, \tag{3.18}$$

$$\mathbf{GADF} = \begin{bmatrix} \sin(\phi_1 - \phi_1) & \sin(\phi_1 - \phi_2) & \ldots & \sin(\phi_1 - \phi_9) \\ \sin(\phi_2 - \phi_1) & \sin(\phi_2 - \phi_2) & \ldots & \sin(\phi_2 - \phi_9) \\ \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_9 - \phi_1) & \sin(\phi_9 - \phi_2) & \ldots & \sin(\phi_9 - \phi_9) \end{bmatrix}. \tag{3.19}$$

Here:

- i and j index the feature dimensions (from 1 to 9);
- **GASF** is a symmetric matrix encoding the correlation strength between feature magnitudes;
- **GADF** is an asymmetric matrix encoding the directional differences between feature magnitudes.

These matrices offer complementary insights into inter-feature dependencies in a form suitable for 2D image modeling.

Step 4: Merging Matrices To integrate both magnitude-based similarity and directional contrast into a unified representation, the **GASF** and **GADF** are fused into a single hybrid matrix $\mathbf{GAF} \in \mathbb{R}^{9\times 9}$ using an asymmetric strategy:

$$\mathrm{GAF}_{ij} = \begin{cases} \mathrm{GASF}_{ij}, & \text{if } i \geq j, \\ \mathrm{GADF}_{ij}, & \text{if } i < j, \end{cases} \quad \text{for } i, j = 1, 2, \ldots, 9. \tag{3.20}$$

This hybrid composition results in:

- The lower triangle ($i \geq j$) encoding self-similarity via angular summation (GASF);
- The upper triangle ($i < j$) capturing inter-feature contrast through angular differences (GADF).

The resulting fused Gramian matrix **GAF** takes the following form:

$$\mathbf{GAF} = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \sin(\phi_1 - \phi_2) & \cdots & \sin(\phi_1 - \phi_9) \\ \cos(\phi_2 + \phi_1) & \cos(\phi_2 + \phi_2) & \cdots & \sin(\phi_2 - \phi_9) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\phi_9 + \phi_1) & \cos(\phi_9 + \phi_2) & \cdots & \cos(\phi_9 + \phi_9) \end{bmatrix}. \tag{3.21}$$

This fused matrix retains both symmetric and antisymmetric patterns and serves as the final diagnostic image representation for downstream processing by deep learning models.

Step 5: Grayscale Mapping Finally, the fused **GAF** is rescaled to the standard 8-bit grayscale pixel range $[0, 255]$, producing an image input suitable for convolutional neural networks:

$$\mathrm{GAF}_{ij}^{\mathrm{image}} = 255 \times \frac{\mathrm{GAF}_{ij} - \min(\mathbf{GAF})}{\max(\mathbf{GAF}) - \min(\mathbf{GAF})}, \quad \text{for } i, j = 1, 2, \ldots, 9. \tag{3.22}$$

This pixel normalization maps the full dynamic range of the fused matrix into the valid 8-bit grayscale space, where 0 corresponds to black and 255 to white. The output matrix is:

$$\mathbf{GAF}^{\mathrm{image}} = \begin{bmatrix} \mathrm{GAF}_{11}^{\mathrm{image}} & \mathrm{GAF}_{12}^{\mathrm{image}} & \cdots & \mathrm{GAF}_{19}^{\mathrm{image}} \\ \mathrm{GAF}_{21}^{\mathrm{image}} & \mathrm{GAF}_{22}^{\mathrm{image}} & \cdots & \mathrm{GAF}_{29}^{\mathrm{image}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{GAF}_{91}^{\mathrm{image}} & \mathrm{GAF}_{92}^{\mathrm{image}} & \cdots & \mathrm{GAF}_{99}^{\mathrm{image}} \end{bmatrix}. \tag{3.23}$$

Each element $\mathrm{GAF}_{ij}^{\mathrm{image}}$ in this matrix, Equation (3.23), represents the intensity value of a pixel located at position (i, j), where $i, j = 1, 2, \ldots, 9$. An example of the resulting 9×9 GAF image is shown in Figure 3.3, illustrating the spatial encoding of feature correlations.

Why GAF is Effective for CNNs CNNs excel at identifying local spatial patterns. The GAF representation transforms inter-feature relationships into spatial patterns where:

- The diagonal encodes self-similarity;
- The lower triangle captures magnitude similarity (GASF);
- The upper triangle captures angular difference (GADF).

These structured patterns allow CNNs to learn fault signatures that reflect not just raw feature values, but how features co-vary and interact—critical for reliable transformer fault classification under domain shifts.

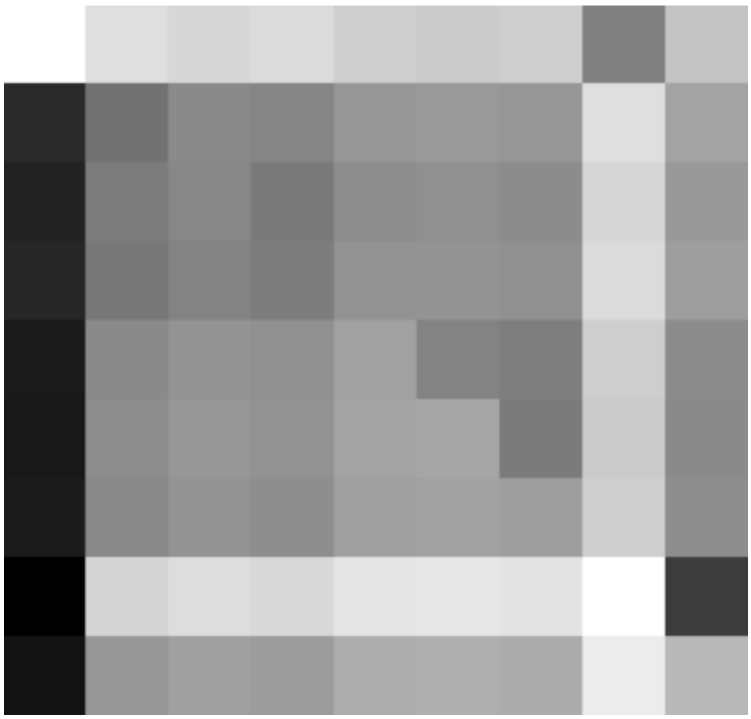

Figure 3.3: Example GAF image obtained after grayscale normalization of the fused GAF matrix. Dark and light regions reflect angular relationships among hybrid diagnostic features.

3.2.2 Data Augmentation

To address the challenge of imbalanced class distribution and limited training data, this study employed data augmentation techniques on the Gramian Angular Field (GAF) images. These augmentations were designed to synthetically increase the diversity of training samples while preserving the underlying class-relevant patterns. Five augmentation methods were applied: Gaussian blur, Gaussian noise, salt-and-pepper noise, brightening, and darkening.

Each augmentation technique introduces a controlled variation to the image, enhancing the model's generalization ability by exposing it to realistic perturbations that could occur in deployment scenarios.

Brightening and Darkening Brightness adjustments were applied to simulate varying lighting conditions or sensor contrast drift, without altering the structural content of the diagnostic image. These transformations operate on the fused GAF image matrix, $\mathbf{GAF}^{\text{image}} \in \mathbb{R}^{9\times 9}$, where each element $\text{GAF}_{ij}^{\text{image}}$ corresponds to the grayscale intensity at pixel location (i, j), for $i, j = 1, 2, \ldots, 9$.

To simulate increased brightness, an arbitrary constant brightening delta, Δb, was added to each pixel:

$$L_{ij}^{+} = \text{GAF}_{ij}^{\text{image}} + \Delta b, \quad \text{for } i, j = 1, 2, \ldots, 9, \quad \Delta b \in \mathbb{R}_{>0} \tag{3.24}$$

where L_{ij}^{+} denotes the pixel intensity after brightening.

To simulate decreased brightness, an arbitrary constant darkening delta, Δd, was subtracted:

$$L_{ij}^{-} = \text{GAF}_{ij}^{\text{image}} - \Delta d, \quad \text{for } i, j = 1, 2, \ldots, 9, \quad \Delta d \in \mathbb{R}_{>0} \tag{3.25}$$

where L_{ij}^{-} represents the pixel intensity after darkening.

In both operations, all modified pixel values were clipped to the valid 8-bit grayscale range $[0, 255]$, ensuring that no overflow or underflow occurs during augmentation. These augmentations help improve the model's robustness to contrast-related variability in real-world deployment.

Salt-and-Pepper Noise Salt-and-pepper noise introduces high-contrast corruption by randomly replacing individual components of the grayscale matrix $\mathbf{GAF}^{\text{image}}$ with extreme intensity values; either black (0) or white (255). For each pixel located at index pair (i, j), the augmented pixel value S_{ij} is computed as:

$$S_{ij} = \begin{cases} 255 & \text{if } r_1 < p \text{ and } r_2 < 0.5, \\ 0 & \text{if } r_1 < p \text{ and } r_2 \geq 0.5, \\ \text{GAF}_{ij}^{\text{image}} & \text{if } r_1 \geq p, \end{cases} \quad \text{for } i, j = 1, 2, \ldots, 9, \tag{3.26}$$

where:

- $p \in [0, 1]$ denotes the noise density (i.e., the probability of a pixel being affected),
- r_1, r_2 are independent random variables sampled from the uniform distribution over $[0, 1]$,
- $\text{GAF}_{ij}^{\text{image}}$ is the original pixel value at location (i, j).

This augmentation mimics impulsive corruption, such as sudden sensor dropout or transmission glitches, enhancing robustness to real-world noise artifacts.

Gaussian Noise Gaussian noise simulates sensor imperfections and environmental disturbances by perturbing each element of the grayscale matrix $\mathbf{GAF}^{\text{image}}$ with additive noise drawn from a zero-mean Gaussian distribution. For each pixel at index pair (i, j), the augmented value N_{ij} is given by:

$$N_{ij} = \text{GAF}_{ij}^{\text{image}} + \mathcal{N}(0, \sigma^2), \quad \text{for } i, j = 1, 2, \ldots, 9, \tag{3.27}$$

where:

- $\mathcal{N}(0, \sigma^2)$ is a normally distributed random variable with mean $\mu = 0$ and variance σ^2,
- $\sigma \in \mathbb{R}_{>0}$ controls the standard deviation and hence the intensity of the noise.

After noise application, all pixel values are clipped to the valid 8-bit grayscale range $[0, 255]$ to prevent overflow and ensure compatibility with image-based models.

Gaussian Blur Gaussian blur is a smoothing operation that attenuates high-frequency components in the image, reducing noise and suppressing small spatial variations. This is particularly useful for simulating low-resolution imaging or camera defocus, enabling the model to generalize better under subtle distortions.

Gaussian Blur Gaussian blur is a smoothing operation that attenuates high-frequency components in the image, reducing noise and suppressing fine spatial details. This transformation is commonly used to simulate low-resolution imaging or minor defocus, enhancing the model's robustness to subtle variations in image sharpness.

Formally, for a given pixel at position (i, j) in the 9×9 grayscale GAF image, $\mathbf{GAF}^{\text{image}}$, the blurred pixel intensity B_{ij} is computed by averaging the intensities of neighboring pixels in a square window centered at (i, j), weighted by a Gaussian function:

$$B_{ij} = \sum_{m=-k}^{k} \sum_{n=-k}^{k} \text{GAF}^{\text{image}}_{i+m\, j+n} \cdot \mathcal{G}(m, n), \quad \text{for } i, j = 1, 2, \ldots, 9, \tag{3.28}$$

where:

- B_{ij} is the blurred pixel intensity at location (i, j),
- $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ are relative offsets within the local window, i.e., they take values in the range $[-k, k]$,
- $k \in \mathbb{N}_{>0}$ defines the half-size of the convolutional kernel (e.g., $k = 1$ gives a 3×3 kernel, $k = 2$ gives a 5×5 kernel),
- $\mathcal{G}(m, n)$ is the Gaussian weight applied to offset (m, n), given by:

$$\mathcal{G}(m, n) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{m^2 + n^2}{2\sigma^2}\right), \tag{3.29}$$

- $\sigma \in \mathbb{R}_{>0}$ is the standard deviation of the Gaussian kernel, controlling the extent of smoothing.

In this formulation, the indices m and n sweep over the neighborhood surrounding the central pixel (i, j), where smaller values of m and n refer to pixels above/left of the center, and larger values refer to pixels below/right. The Gaussian kernel $\mathcal{G}(m, n)$

ensures that pixels closer to the center have higher influence on the output B_{ij}, while more distant pixels contribute less.

Convolving this kernel over the image produces a softened version of $\mathbf{GAF}^{\text{image}}$, where local intensity variations are smoothed out while preserving broader structural patterns. The final output pixel values are clipped to the standard grayscale range $[0, 255]$ to ensure valid image representation.

Visualization of Augmented Samples Figure 3.4 shows a representative original GAF image and its augmented versions. Each transformation preserves the overall structure of the image while introducing a distinct variation.

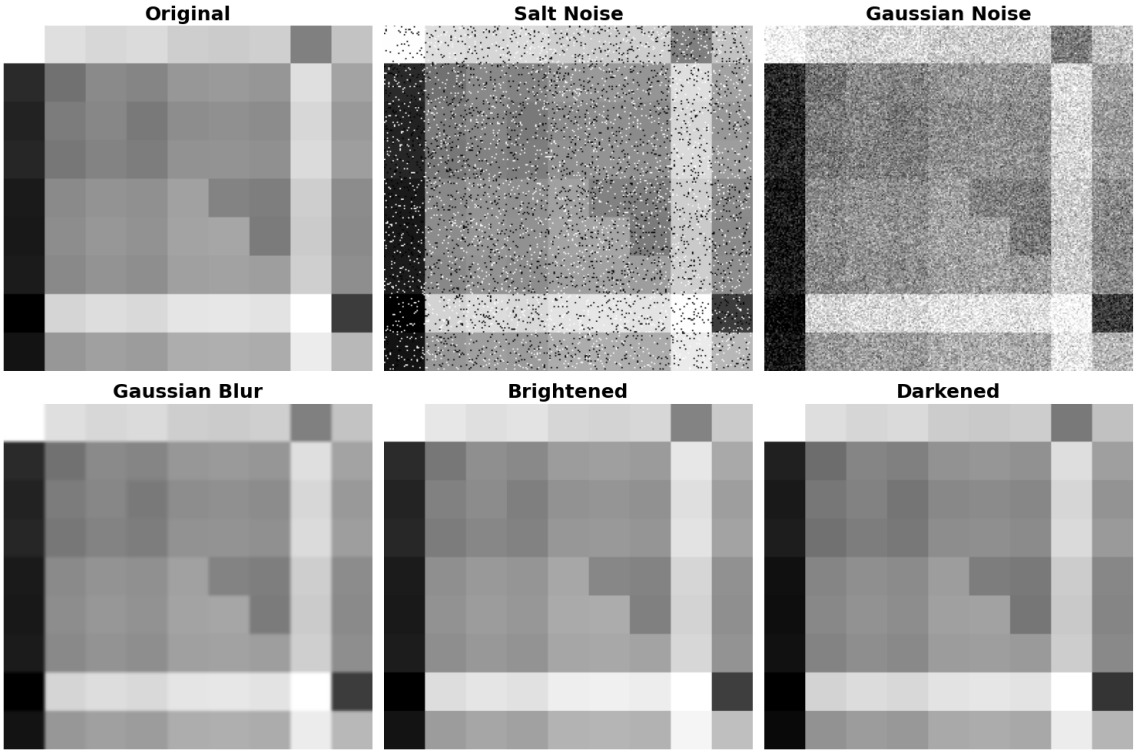

Figure 3.4: Examples of data augmentation techniques applied to a GAF image. Variants include Gaussian noise, salt-and-pepper noise, blur, brightening, and darkening.

3.2.3 Model Architecture

Convolutional Neural Networks (CNNs) are a class of deep learning models designed to process structured input data such as two-dimensional images. Originating from neuroscience-inspired architectures like LeNet-5 [18] and later popularized by deeper models such as AlexNet [16], CNNs have become the standard for image-based classification tasks. In this study, a CNN is employed to classify grayscale Gramian Angular Field (GAF) images, denoted as $\mathbf{GAF}^{\text{image}}$, which are constructed from hybrid DGA ratio features for transformer fault diagnosis.

The proposed CNN architecture is composed of two main stages: a feature extraction module and a classification head.

The feature extraction stage begins with a convolutional layer that applies 32 filters to the input image, $\mathbf{GAF}^{\text{image}}$, learning spatial patterns such as angular trends or localized intensities that are characteristic of different fault classes. Each filter generates a corresponding feature map, and a ReLU activation is applied to introduce non-linearity by retaining only positive responses. A 2×2 max pooling operation follows, which downsamples each feature map by selecting the highest activation within non-overlapping windows. This process helps the network become invariant to small spatial shifts while reducing computational complexity.

A second convolutional layer then applies 64 filters to the pooled outputs from the previous layer. As before, ReLU activation and 2×2 max pooling are applied in sequence. This deepens the network's ability to detect higher-order spatial patterns by hierarchically building upon the localized features extracted in earlier layers.

The output of the final pooling layer is a stack of two-dimensional feature maps, which is then flattened into a one-dimensional feature vector. This vector is passed to a fully connected layer that transforms it into a compact representation for classification. To improve generalization and reduce the risk of overfitting, dropout is applied during training by randomly deactivating a subset of neurons in the fully connected layer.

The final output layer produces a vector of logits, which is then passed through a softmax activation function to yield normalized class probabilities. Each probability

corresponds to one of the five fault categories defined in the classification task.

Overall, this compact CNN architecture efficiently captures both low-level and high-level patterns in the input, $\mathbf{GAF}^{\text{image}}$, while maintaining strong classification performance. An overview of the architecture is shown in Figure 3.5.

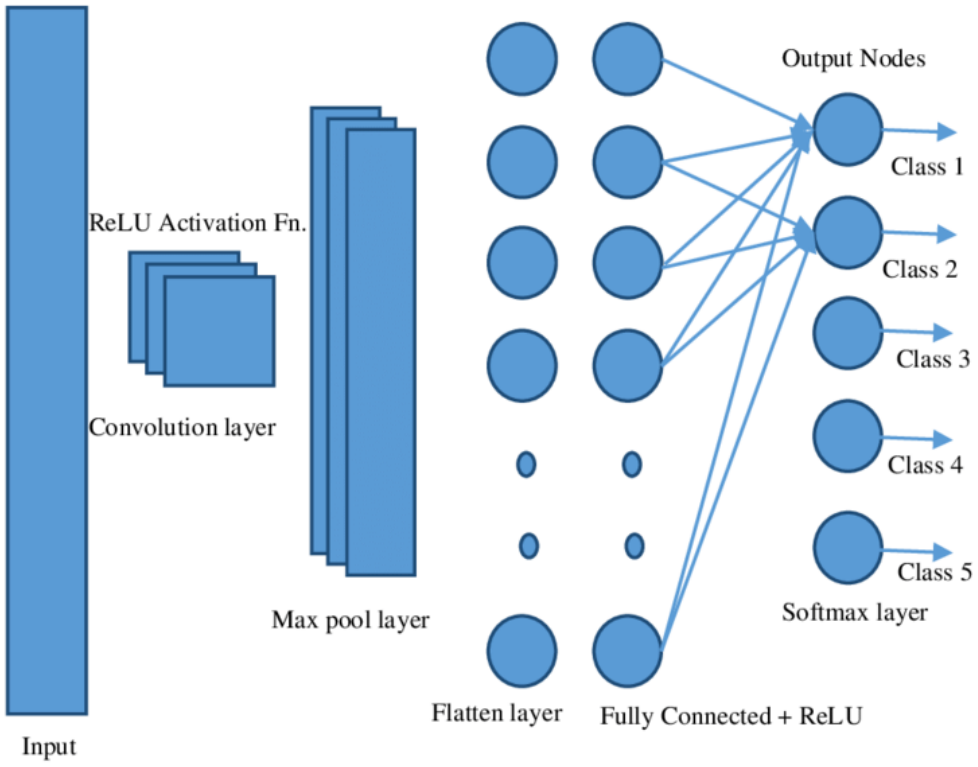

Figure 3.5: Architecture of the proposed CNN for fault classification from $\mathbf{GAF}^{\text{image}}$.

3.2.4 Foundations of Weighted Domain Adaptation

In domain adaptation, the core objective is to align the distributions of the source domain $\mathcal{D}_S$ and the target domain $\mathcal{D}_T$ so that a model trained on labeled source data generalizes well to unlabeled or sparsely labeled target data. This section introduces the foundational components of our feature-weighted domain adaptation strategy: the Kolmogorov–Smirnov (K-S) statistic, Maximum Mean Discrepancy (MMD), and

Correlation Alignment (CORAL). Together, these tools facilitate both distributional and structural alignment, and their integration ensures that features with significant domain shift receive proportionally greater emphasis during training.

Kolmogorov–Smirnov Statistic

The Kolmogorov–Smirnov (K-S) test is a classical non-parametric method introduced by Kolmogorov and Smirnov [23] to compare the distributions of two datasets without assuming any underlying distributional form. In our domain adaptation framework, it serves as a statistical tool for quantifying the distributional shift between corresponding features in the source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$.

The input to our model is the hybrid diagnostic feature vector, Equation (3.11), which captures both the relative concentrations and the logarithmic relationships among dissolved gases. To assess how each component g_i varies across domains, we compute a separate K-S statistic for each feature.

Let $\{g_{i1}^{\mathcal{D}_S}, g_{i2}^{\mathcal{D}_S}, \ldots, g_{in}^{\mathcal{D}_S}\}$ and $\{g_{i1}^{\mathcal{D}_T}, g_{i2}^{\mathcal{D}_T}, \ldots, g_{im}^{\mathcal{D}_T}\}$ denote the values of the feature g_i across all n source samples and m target samples, respectively. The empirical cumulative distribution functions (CDFs) of the source and target domains for this feature are given by:

- $F_i(x) = \frac{1}{n}\sum_{w=1}^{n} \mathbf{1}_{(-\infty, x]}\left(g_{iw}^{\mathcal{D}_S}\right)$: the proportion of source-domain values of the feature g_i with n number of samples less than or equal to a threshold x,
- $G_i(x) = \frac{1}{m}\sum_{z=1}^{m} \mathbf{1}_{(-\infty, x]}\left(g_{iz}^{\mathcal{D}_T}\right)$: the proportion of target-domain values of the feature g_i with m number of samples less than or equal to a threshold x,
- $x \in \mathbb{R}$: the evaluation threshold for the cumulative functions,
- $\mathbf{1}_{(-\infty, x]}(\ )$: indicator function, which returns 1 if its argument is less than or equal to x, and 0 otherwise.

The K-S statistic for feature g_i is then:

$$D_i = \sup_{x \in \mathbb{R}} \left| F_i(x) - G_i(x) \right|, \tag{3.30}$$

where sup denotes the supremum, or least upper bound. For empirical (finite-sample) distributions, this supremum is equivalent to the maximum over all observed x values:

$$D_i = \max_x |F_i(x) - G_i(x)| . \tag{3.31}$$

This scalar value D_i reflects the largest observed discrepancy between the empirical CDFs of the source and target domains for the i^{th} feature. A higher D_i implies a greater degree of domain shift for that particular diagnostic dimension.

As visualized in Figure 3.6, the K-S statistic corresponds to the largest vertical gap between the two empirical CDF curves.

Use in Our Framework We apply the K-S test to each of the nine hybrid features $g_1, \ldots, g_9$ to obtain a vector of divergence scores:

$$\boldsymbol{D} = [D_1, D_2, \ldots, D_9]^\top. \tag{3.32}$$

This vector is then normalized and incorporated into the domain adaptation loss as feature-wise importance weights. Features with larger D_i values are emphasized more heavily during alignment, allowing the model to prioritize correction of the most misaligned components of the feature space. This weighting mechanism ensures the adaptation process focuses on reducing the most impactful domain shifts.

Maximum Mean Discrepancy (MMD)

Maximum Mean Discrepancy (MMD), first introduced by Gretton et al. [7], is a non-parametric statistical measure for quantifying the distance between two probability distributions P and Q, based on samples drawn from each. In our framework, P and Q correspond to the feature distributions in the source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, respectively.

The fundamental principle of MMD is to map samples from both domains into a Reproducing Kernel Hilbert Space (RKHS) via a kernel function $k(\cdot, \cdot)$, and to measure the distance between their mean embeddings. A smaller distance implies

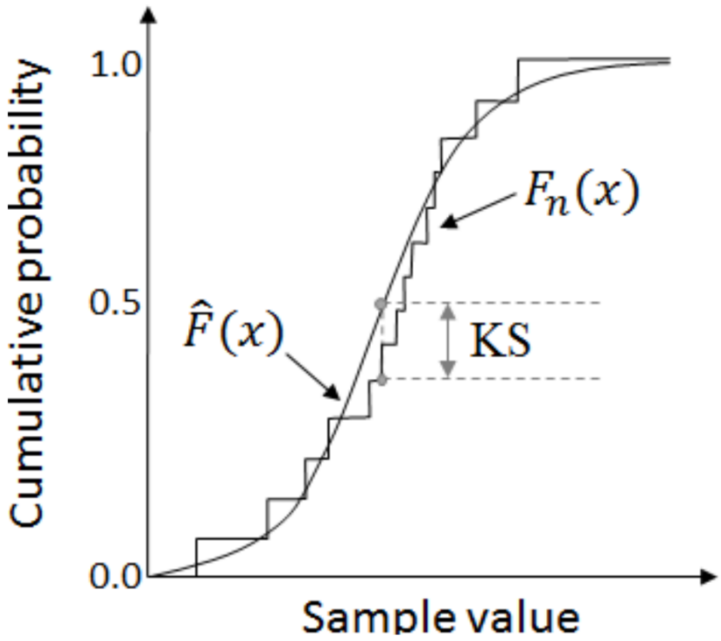

Figure 3.6: Visual demonstration of the Kolmogorov–Smirnov statistic as the maximum vertical distance between two empirical CDFs. The stair-stepped curve denotes the source-domain empirical CDF, while the smooth curve represents the target-domain CDF. The vertical gap labeled "KS" reflects the statistic D_i for a given feature [44].

better domain alignment, while a larger distance indicates stronger distributional shift.

Formally, the squared MMD is computed as:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_S},\boldsymbol{g}^{\mathcal{D}_{S'}}}[k(\boldsymbol{g}^{\mathcal{D}_S},\boldsymbol{g}^{\mathcal{D}_{S'}})]+\mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_T},\boldsymbol{g}^{\mathcal{D}_{T'}}}[k(\boldsymbol{g}^{\mathcal{D}_T},\boldsymbol{g}^{\mathcal{D}_{T'}})]-2\mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_S},\boldsymbol{g}^{\mathcal{D}_T}}[k(\boldsymbol{g}^{\mathcal{D}_S},\boldsymbol{g}^{\mathcal{D}_T})], \tag{3.33}$$

where:

- $\boldsymbol{g}^{\mathcal{D}_S}, \boldsymbol{g}^{\mathcal{D}_{s'}}$ are two independently sampled hybrid feature vectors from the source domain $\mathcal{D}_S$ with feature distribution P, each of the form $\boldsymbol{g} = [g_1, g_2, \ldots, g_9]^\top$,
- $\boldsymbol{g}^{\mathcal{D}_T}, \boldsymbol{g}^{\mathcal{D}_{T'}}$ are two independently sampled hybrid feature vectors from the target domain $\mathcal{D}_T$ with feature distribution Q, each of the form $\boldsymbol{g} = [g_1, g_2, \ldots, g_9]^\top$,

- $k(\cdot, \cdot)$ is a positive-definite kernel function that measures similarity between two feature vectors.

The three terms in Equation (3.33) can be interpreted as follows:

- The first term, $\mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_S}, \boldsymbol{g}^{\mathcal{D}_{S'}}}[k(\boldsymbol{g}^{\mathcal{D}_S}, \boldsymbol{g}^{\mathcal{D}_{S'}})]$, computes the expected similarity between two source-domain samples. A high value indicates that source features are tightly clustered in the kernel space.

- The second term, $\mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_T}, \boldsymbol{g}^{\mathcal{D}_{T'}}}[k(\boldsymbol{g}^{\mathcal{D}_T}, \boldsymbol{g}^{\mathcal{D}_{T'}})]$, computes the expected similarity between two target-domain samples. A high value indicates that target features are tightly clustered in the kernel space.

- The third term, $\mathbb{E}_{\boldsymbol{g}^{\mathcal{D}_S}, \boldsymbol{g}^{\mathcal{D}_T}}[k(\boldsymbol{g}^{\mathcal{D}_S}, \boldsymbol{g}^{\mathcal{D}_T})]$, captures the expected cross-domain similarity between source and target feature vectors.

Together, these terms quantify the discrepancy between the mean embeddings of the two domains in kernel space. A large MMD value implies that the distributions of source and target features differ substantially. Minimizing this loss during training promotes the alignment of latent representations across domains, encouraging domain-invariant feature learning.

In this work, we adopt the Gaussian Radial Basis Function (RBF) kernel:

$$k(\boldsymbol{g}, \boldsymbol{g}') = \exp\left(-\frac{\|\boldsymbol{g} - \boldsymbol{g}'\|^2}{2\sigma^2}\right), \tag{3.34}$$

where $\sigma > 0$ is the kernel bandwidth that controls the sensitivity of the similarity measure.

Intuition As illustrated in Figure 3.7, the hybrid feature vectors from the source and target domains are first projected into a shared RKHS space using the kernel function. MMD then computes the distance between their mean representations. A higher value implies that the domains are misaligned, whereas minimizing this discrepancy encourages the network to learn domain-invariant features.

Advantages for Deep Learning MMD is particularly attractive for deep transfer learning for the following reasons:

- **Non-parametric:** MMD makes no assumptions about the underlying distribution shape, making it robust to real-world data irregularities.
- **Kernel-based:** Through kernels, it can capture both linear and nonlinear similarities.
- **Differentiable:** It can be integrated into the loss function and optimized via gradient-based methods in deep networks.

Use in Our Framework Following Long et al. [22], we incorporate MMD as an auxiliary loss to guide domain adaptation:

$$\mathcal{L}_{\text{MMD}} = \text{MMD}^2(P, Q). \tag{3.35}$$

Specifically, we compute the MMD loss between the source and target feature representations (i.e., $\boldsymbol{g}^{\mathcal{D}_S}$ and $\boldsymbol{g}^{\mathcal{D}_T}$) at intermediate layers of the CNN. This encourages alignment of the learned features across domains, mitigating distributional shifts and improving generalization to unseen target samples.

Correlation Alignment (CORAL)

Correlation Alignment (CORAL), proposed by Sun et al. [31], is a domain adaptation technique that aligns the second-order statistics (i.e., covariances) of source and target feature distributions. Unlike Maximum Mean Discrepancy (MMD), which compares full probability distributions in a reproducing kernel Hilbert space (RKHS), CORAL minimizes the discrepancy between empirical covariance matrices derived from latent features in each domain.

Motivation and Intuition Distribution shift between domains often manifests not only in the mean values of features, but also in how these features vary together. CORAL aims to reduce this shift by aligning the internal feature structures,

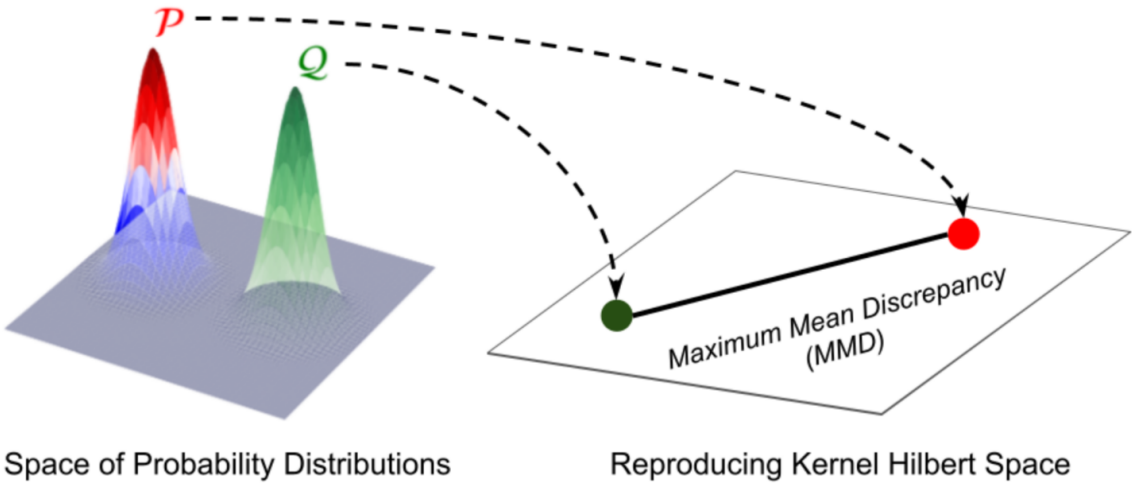

Figure 3.7: Conceptual illustration of MMD in latent space [24]. Source and target features are mapped via a kernel function, and the distance between their mean embeddings is minimized.

specifically the covariance relationships, across the source and target domains. As illustrated in Figure 3.8, aligning second-order statistics reshapes the target distribution to mimic the source structure, facilitating generalization without access to target labels.

CORAL Loss Definition Let $\boldsymbol{g}^{\mathcal{D}_S} \in \mathbb{R}^{n\times 9}$ and $\boldsymbol{g}^{\mathcal{D}_T} \in \mathbb{R}^{m\times 9}$ denote matrices of 9-dimensional hybrid feature vectors extracted from the source and target domains, respectively. The CORAL loss is defined as:

$$\mathcal{L}_{\text{CORAL}} = \frac{1}{4d^2} \left\| \boldsymbol{C}^{\mathcal{D}_S} - \boldsymbol{C}^{\mathcal{D}_T} \right\|_F^2, \tag{3.36}$$

where:

- $\boldsymbol{C}^{\mathcal{D}_S} \in \mathbb{R}^{d\times d}$: empirical covariance matrix of source features $\boldsymbol{g}^{\mathcal{D}_S}$,
- $\boldsymbol{C}^{\mathcal{D}_T} \in \mathbb{R}^{d\times d}$: empirical covariance matrix of target features $\boldsymbol{g}^{\mathcal{D}_T}$,
- d: dimensionality of the feature space (here, $d = 9$ for the hybrid diagnostic vector),

- $\|\cdot\|_F^2$: squared matrix Frobenius norm, measuring element-wise squared differences between matrices:

$$\|A - B\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{d}(a_{ij} - b_{ij})^2. \tag{3.37}$$

Covariance Matrix Computation Given a matrix of features $X \in \mathbb{R}^{n\times d}$, where each row is a hybrid vector $\boldsymbol{g}$, the empirical covariance matrix $C \in \mathbb{R}^{d\times d}$ is computed element-wise as:

$$C_{pq} = \frac{1}{n-1}\sum_{i=1}^{n}\left(g_{pi} - \mu_p\right)\left(g_{qi} - \mu_q\right), \tag{3.38}$$

where:

- g_{pi}: the p^{th} feature of the i^{th} sample,
- g_{qi}: the q^{th} feature of the i^{th} sample,
- $\mu_p = \frac{1}{n}\sum_{i=1}^{n} g_{pi}$: the mean of the p^{th} feature across all samples,
- $\mu_q = \frac{1}{n}\sum_{i=1}^{n} g_{qi}$: the mean of the q^{th} feature across all samples,
- C_{pq}: the covariance between feature p and feature q.

The full covariance matrix $\boldsymbol{C} \in \mathbb{R}^{9\times 9}$ is then:

$$\boldsymbol{C} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{19} \\ C_{21} & C_{22} & \cdots & C_{29} \\ \vdots & \vdots & \ddots & \vdots \\ C_{91} & C_{92} & \cdots & C_{99} \end{bmatrix}, \tag{3.39}$$

This symmetric matrix captures the second-order (co-variation) relationships among the hybrid diagnostic features across the sample size n.

Interpretation Minimizing the CORAL loss encourages the network to learn a feature space where the covariance structures of the source and target domains are similar. This form of alignment promotes a domain-invariant representation, improving transfer performance under covariate shift.

Advantages of CORAL CORAL is particularly attractive in deep transfer learning settings due to:

- **No kernel or hyperparameters:** Unlike MMD, CORAL does not require choosing a kernel function or tuning bandwidths.
- **Low computational overhead:** Involves only matrix operations, making it efficient even for high-dimensional features.
- **Interpretability:** The alignment of second-order statistics provides a clear geometric and statistical interpretation.
- **Gradient-friendly:** The loss is fully differentiable and easily integrated into neural network training.

Use in Our Framework Following Sun et al. [31], we apply CORAL as an auxiliary loss term at intermediate feature layers of our CNN. Specifically, we compute the CORAL loss between the batchwise source and target hybrid feature matrices $\boldsymbol{g}^{\mathcal{D}_S}$ and $\boldsymbol{g}^{\mathcal{D}_T}$, encouraging the model to align the statistical structure of the two domains in the learned representation space. This promotes better generalization to unseen target data and complements the mean-based alignment performed by MMD.

MMD + CORAL Hybridization

Recent research has shown that combining MMD and CORAL provides complementary benefits in domain adaptation. MMD excels at aligning global distributional properties by matching the means in a reproducing kernel Hilbert space (RKHS), while CORAL focuses on matching the second-order statistics (i.e., covariances), thus capturing internal feature correlations.

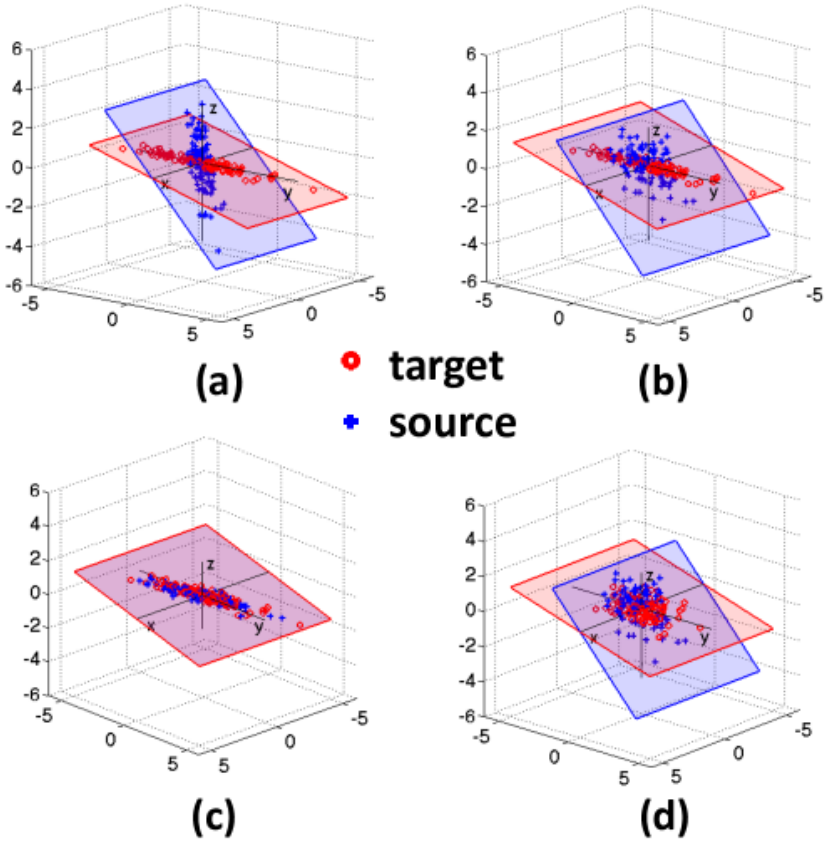

Figure 3.8: Conceptual illustration of CORAL: source and target features are projected into a shared latent space where their covariance structures are aligned [30].

To leverage both perspectives, a hybrid domain alignment loss is often constructed as a weighted sum of the individual MMD and CORAL losses:

$$\mathcal{L}_{\text{domain}} = \beta \cdot \mathcal{L}_{\text{MMD}} + (1 - \beta) \cdot \mathcal{L}_{\text{CORAL}}, \tag{3.40}$$

where:

- $\mathcal{L}_{\text{MMD}}$ measures the mean discrepancy between source and target distributions, Equation (3.35),

- $\mathcal{L}_{\text{CORAL}}$ aligns feature covariances, Equation (3.36),
- $\beta \in [0, 1]$ is a hyperparameter that balances the emphasis between the two losses:
  - If $\beta = 1$, only MMD is used.
  - If $\beta = 0$, only CORAL is applied.
  - Intermediate values combine both in proportion.

In our proposed method, we further enhance this hybrid loss using a feature-level weighting scheme derived from the Kolmogorov–Smirnov (K-S) statistics. Instead of treating all features equally, we modulate the domain loss based on the degree of distribution shift observed in each feature dimension. Features exhibiting higher K-S values (i.e., more discrepancy between source and target) are given higher importance during adaptation, improving alignment in the most divergent dimensions and leading to more effective transfer learning.

This weighted hybridization allows the framework to not only benefit from the strengths of both MMD and CORAL but also adaptively prioritize alignment efforts based on real statistical evidence of domain shift.

3.2.5 Weighted Domain Adaptation Using K–S Statistics

To enhance the effectiveness of domain adaptation in transformer fault diagnosis, we propose a feature-weighted strategy that leverages the Kolmogorov–Smirnov (K–S) statistic to quantify domain shift. Unlike conventional approaches such as plain MMD or CORAL that treat all feature dimensions equally, our method up-weights features that exhibit stronger cross-domain distributional differences. This selective emphasis allows the model to focus alignment pressure on the parts of the feature space where it matters most.

Motivation: From Drift Detection to Alignment Guidance In Dissolved Gas Analysis (DGA), diagnostic insight often emerges not from isolated features,

but from their pairwise relationships and ratios. Thus, we not only assess individual feature shift using K–S statistics, but also encode these shifts into a pairwise structure that reflects joint importance. The goal is to bridge classical two-sample statistical testing with modern representation learning, forming a data-driven alignment objective that adapts to the real distribution gap.

Weight Derivation via GAF-Encoded K–S Scores Let the hybrid feature vectors from the source and target domains be denoted as $\boldsymbol{g}^{\mathcal{D}_S} \sim P$ and $\boldsymbol{g}^{\mathcal{D}_T} \sim Q$, respectively. For each feature g_i, we compute the K–S statistic between its empirical distributions in $\mathcal{D}_S$ and $\mathcal{D}_T$. These scalar values are normalized and encoded into a 2D Gramian Angular Field (GAF) image to capture angular dependencies across features.

This process produces a symmetric matrix $\boldsymbol{W} \in \mathbb{R}^{9\times 9}$, where each element W_{ij} reflects the pairwise alignment importance between features g_i and g_j. To prevent vanishing gradients during training, the final weight values are normalized to the $[0, 1]$ range and stabilized using a small constant $\epsilon > 0$.

Integration into MMD and CORAL The weight matrix $\boldsymbol{W}$ is incorporated into the MMD and CORAL losses to emphasize high-shift feature pairs.

Weighted MMD To implement this formulation in practice, the expectations in Equation (3.33) are approximated empirically using mini-batch samples and weighted kernel evaluations as follows:

$$\begin{aligned}\mathcal{L}_{\text{MMD}}^{weighted} =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\cdot k(\boldsymbol{g}_i^{\mathcal{D}_S}, \boldsymbol{g}_j^{\mathcal{D}_S}) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} W_{ij}\cdot k(\boldsymbol{g}_i^{\mathcal{D}_T}, \boldsymbol{g}_j^{\mathcal{D}_T}) \\ & - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\cdot k(\boldsymbol{g}_i^{\mathcal{D}_S}, \boldsymbol{g}_j^{\mathcal{D}_T}),\end{aligned} \tag{3.41}$$

.

where:

- n, m: Number of samples in the source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, respectively.
- $\boldsymbol{g}_i^{\mathcal{D}_S}, \boldsymbol{g}_j^{\mathcal{D}_S} \in \mathbb{R}^9$: Hybrid feature vectors from the source domain.
- $\boldsymbol{g}_i^{\mathcal{D}_T}, \boldsymbol{g}_j^{\mathcal{D}_T} \in \mathbb{R}^9$: Hybrid feature vectors from the target domain.
- $W_{ij} \in (0, 1]$: Feature-pair importance weights derived from the GAF-transformed K–S statistics. These values emphasize alignment along feature directions with stronger domain shift.
- $k(\cdot, \cdot)$: Gaussian Radial Basis Function (RBF) kernel that maps feature vectors into a Reproducing Kernel Hilbert Space (RKHS) and computes similarity.

Weighted CORAL For covariance alignment, each element is scaled using W_{pq}:

$$C_{pq}^{\mathcal{D}_S} = \frac{1}{n-1} W_{pq} \sum_{i=1}^{n} \left(g_{pi}^{\mathcal{D}_S} - \mu_p^{\mathcal{D}_S}\right) \left(g_{qi}^{\mathcal{D}_S} - \mu_q^{\mathcal{D}_S}\right), \tag{3.42}$$

$$C_{pq}^{\mathcal{D}_T} = \frac{1}{m-1} W_{pq} \sum_{i=1}^{m} \left(g_{pi}^{\mathcal{D}_T} - \mu_p^{\mathcal{D}_T}\right) \left(g_{qi}^{\mathcal{D}_T} - \mu_q^{\mathcal{D}_T}\right), \tag{3.43}$$

$$\mathcal{L}_{\text{CORAL}}^{weighted} = \frac{1}{4d^2} \sum_{p=1}^{d} \sum_{q=1}^{d} \left(C_{pq}^{\mathcal{D}_s} - C_{pq}^{\mathcal{D}_T}\right)^2, \tag{3.44}$$

where:

- d: Dimensionality of the hybrid feature vector $\boldsymbol{g} \in \mathbb{R}^d$ ($d = 9$).
- n, m: Number of samples in the source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, respectively.
- $g_{pi}^{\mathcal{D}_S}$, $g_{qi}^{\mathcal{D}_S}$: The p^{th} and q^{th} feature components of the i^{th} sample from the source domain.
- $g_{pi}^{\mathcal{D}_T}$, $g_{qi}^{\mathcal{D}_T}$: The p^{th} and q^{th} feature components of the i^{th} sample from the target domain.

- $\mu_p^{\mathcal{D}_S}, \mu_q^{\mathcal{D}_S}$: Mean values of the p^{th} and q^{th} features across all source samples.
- $\mu_p^{\mathcal{D}_T}, \mu_q^{\mathcal{D}_T}$: Mean values of the p^{th} and q^{th} features across all target samples.
- $W_{pq} \in (0, 1]$: Pairwise feature alignment weight computed from the GAF-transformed K–S statistics, reflecting the significance of aligning covariance between feature pair (p, q).
- $C_{pq}^{\mathcal{D}_S}, C_{pq}^{\mathcal{D}_T}$: Weighted covariance between features p and q in the source and target domains, respectively.
- $\mathcal{L}_{\text{CORAL}}^{\text{weighted}}$: Final loss measuring the weighted Frobenius distance between source and target covariance matrices.

Total Loss Function The final objective combines classification and domain alignment losses:

$$\mathcal{L}_{\text{total}}^{weighted} = \alpha \cdot \mathcal{L}_{\text{classification}} + (1 - \alpha) \cdot \mathcal{L}_{\text{domain}}^{weighted}, \tag{3.45}$$

$$\mathcal{L}_{\text{domain}}^{weighted} = \beta \cdot \mathcal{L}_{\text{MMD}}^{weighted} + (1 - \beta) \cdot \mathcal{L}_{\text{CORAL}}^{weighted}, \tag{3.46}$$

where:

- $\mathcal{L}_{\text{classification}}$: Cross-entropy loss computed over labeled source domain samples, guiding supervised learning.
- $\mathcal{L}_{\text{MMD}}^{\text{weighted}}$: Feature-weighted Maximum Mean Discrepancy loss, aligning marginal distributions across domains.
- $\mathcal{L}_{\text{CORAL}}^{\text{weighted}}$: Feature-weighted CORAL loss, aligning second-order statistics (covariances) between domains.
- $\mathcal{L}_{\text{domain}}^{\text{weighted}}$: Combined feature-weighted domain alignment loss.
- $\mathcal{L}_{\text{total}}^{\text{weighted}}$: Total training objective that balances classification and the feature-weighted domain alignment losses.

- $\alpha \in [0, 1]$: Hyperparameter controlling the relative importance of classification vs. domain alignment objectives.
- $\beta \in [0, 1]$: Hyperparameter controlling the balance between MMD and CORAL components within the Feadomain alignment term.

Advantages and Novelty

1. **K–S attention bridge:** First known application of K–S statistics as a guidance signal for feature-pair alignment in domain adaptation.
2. **Pairwise weighting:** GAF-encoded weights capture fine-grained dependencies between hybrid diagnostic features.
3. **Model-agnostic design:** Weighting layer is modular and can be applied to any MMD- or CORAL-based framework.
4. **Adaptive regularization:** Shifts alignment pressure to domain-sensitive subspaces, reducing over-regularization of already aligned features.

Chapter 4

Results

4.1 Overview

This chapter presents a comprehensive evaluation of the proposed feature-weighted domain adaptation method for power transformer fault diagnosis using Dissolved Gas Analysis (DGA) data. The evaluation compares three approaches: Fine-Tuning, standard domain adaptation using MMD and CORAL (MC), and the proposed method that enhances MC by incorporating feature-specific weights based on the Kolmogorov–Smirnov (K-S) statistics (MCW). The comparison is carried out using a range of metrics including classification accuracy, F1-score, confusion matrices, and distribution shift metrics such as the Average Kullback–Leibler Divergence (AKLD) and pixel intensity histograms of Gramian Angular Field (GAF) images. Additionally, the sensitivity of the methods to varying training sample sizes and hyperparameters is examined in detail.

4.2 Evaluation Metrics

This section introduces and defines the evaluation metrics used to assess model performance in power transformer fault diagnosis. Two major categories of metrics are considered: (1) classification performance metrics used to evaluate the accuracy and

effectiveness of predictions, and (2) domain discrepancy metrics used to measure the statistical differences between the source and target datasets. The metrics selected are essential to interpret the results discussed in this chapter.

4.2.1 Classification Performance Metrics

Accuracy

Accuracy is a widely used metric that represents the proportion of correct predictions made by the model over the total number of predictions. It provides a simple indication of the overall effectiveness of a classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Despite its intuitive appeal, accuracy may be misleading in imbalanced datasets, such as those encountered in transformer fault diagnosis, where certain fault types are underrepresented.

Precision, Recall, and F1-Score

To address the limitations of accuracy, especially in the presence of class imbalance, precision, recall, and F1-score provide more nuanced insights:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Precision reflects the correctness among positive predictions, while recall indicates the model's ability to detect actual positives. The F1-score balances both metrics and is especially useful in multiclass classification tasks involving imbalanced data, as is common in power transformer datasets.

Confusion Matrix

A confusion matrix provides a detailed breakdown of classification results by comparing true class labels with predicted ones. For a multiclass classification problem with K classes, the confusion matrix is a $K \times K$ matrix where the element at row i and column j represents the number of samples from class i predicted as class j.

In this study, a normalized row-wise confusion matrix is used to better visualize class-wise performance. Each row corresponds to the true class, and the values indicate the percentage distribution of predictions across classes. This format highlights which classes are most frequently confused by the model.

4.2.2 Domain Discrepancy Metrics

Average Kullback–Leibler Divergence (AKLD)

The Kullback–Leibler (KL) divergence is a measure of how one probability distribution diverges from a second, expected distribution. For discrete distributions P and Q, the KL divergence is defined as:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{4.5}$$

where:

- $P(i)$: The probability of event i under the true (or reference) distribution P,
- $Q(i)$: The probability of event i under the approximate or learned distribution Q,
- $D_{KL}(P \parallel Q)$: The Kullback–Leibler divergence, quantifying how much information is lost when Q is used to approximate P.

In this study, we compute the Average KL Divergence (AKLD) across all nine features used in the hybrid DGA ratio set to quantify the overall feature-level distribution shift between the source and target datasets. Higher AKLD values indicate

greater divergence. This metric is sensitive to mismatches between probability densities and complements other distribution-based evaluations such as the K-S test.

Pixel Intensity Distribution

To capture visual and structural differences between the domains after applying the Gramian Angular Field (GAF) transformation, pixel intensity distributions are compared. Each GAF image encodes feature relationships into a grayscale matrix, and the histogram of pixel intensities reflects the overall energy and contrast embedded in the data.

4.2.3 Summary

The combination of performance and distribution-based metrics ensures a thorough evaluation of the proposed model. While accuracy and F1-score assess predictive success, AKLD and pixel intensity comparisons provide insight into the statistical and structural alignment of the domains. These metrics will be applied consistently across all experiments in this chapter.

4.3 Case Study and Dataset Overview

In this research, we utilized two distinct datasets. The source dataset, derived from the literature [33], includes samples collected from the Egyptian Electrical Utility [1] and the Indian Utility in the TIFAC laboratory [35], comprising 384 samples. The target dataset, obtained from the IEC TC 10 database, includes 99 samples. Both datasets consist of five fault types: Partial Discharge (PD), Low Energy Discharge (D1), High Energy Discharge (D2), Low and Medium Thermal Fault (T1&T2), and High Thermal Fault (T3).

Figures 4.1 and 4.2 present the label distributions of the source and target datasets, respectively.

Figure 4.3 shows the distribution of labels in the source dataset after augmentation, illustrating how underrepresented classes were balanced through data augmen-

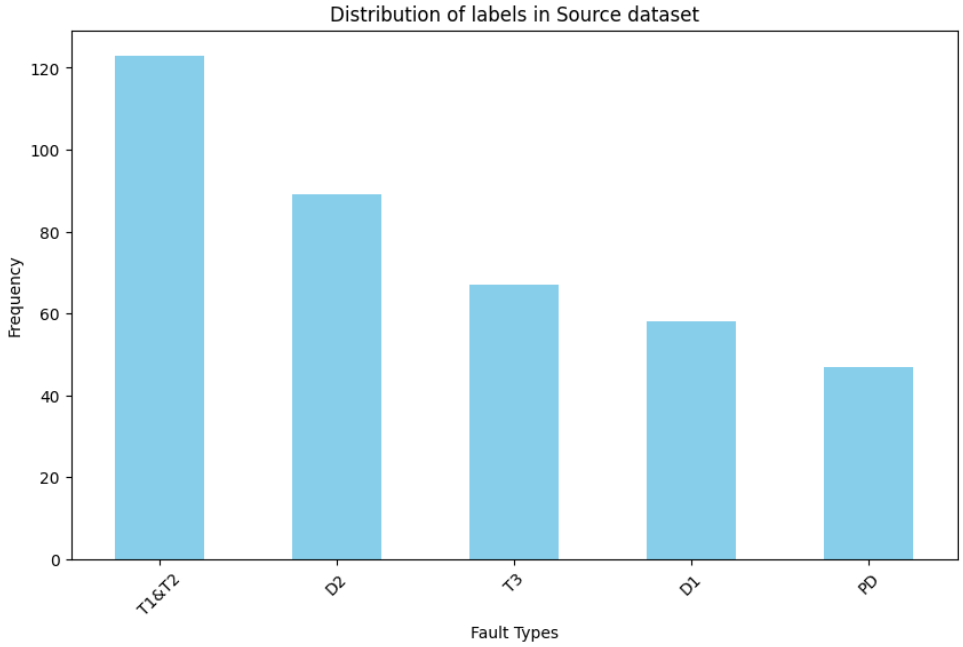

Figure 4.1: Source dataset label distribution.

tation techniques.

To quantify the domain shift, we employed two complementary strategies:

- **Average Kullback–Leibler Divergence (AKLD)**: Calculated across nine hybrid DGA features, resulting in an AKLD of 0.698. This value indicates a significant statistical divergence between the feature distributions of the source and target datasets.

- **Pixel Intensity Distributions of GAF Images**: To capture visual differences in feature structure between domains, we compared the pixel intensity histograms of the GAF images for both the source and target datasets. As shown in Figure 4.4, the source domain images exhibit more extreme values concentrated around the upper and lower bounds of the intensity scale, which corresponds to a highly structured representation. In contrast, the target domain images show a more uniform and flatter intensity distribution, indicating less distinctive patterns or different underlying feature correlations post-GAF

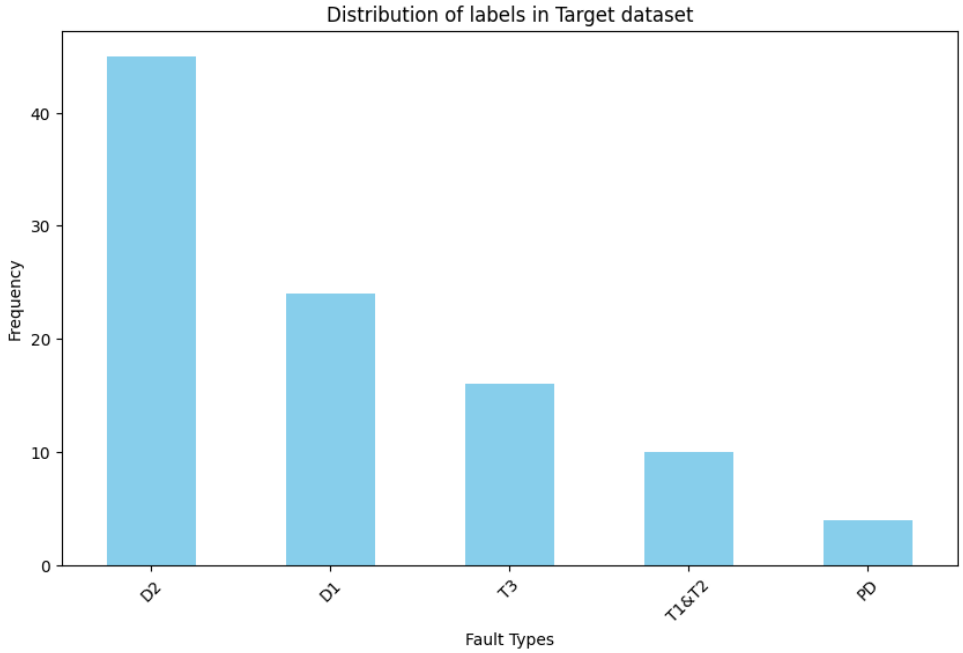

Figure 4.2: Target dataset label distribution.

transformation. These observed differences support the need for domain adaptation methods, as they illustrate how structurally different the transformed representations are, even when derived from similar diagnostic features.

4.4 Correlation Between Features and Fault Types

To investigate the discriminative power of the selected hybrid DGA features, we computed the Pearson correlation coefficients between each feature and the five fault types. Figure 4.5 presents the resulting correlation matrix.

As illustrated in the matrix, several features exhibit moderate to strong correlations with specific fault types, reinforcing their importance in the classification task:

- **C2H4_Ratio** shows a high positive correlation with fault type T3 (0.72),

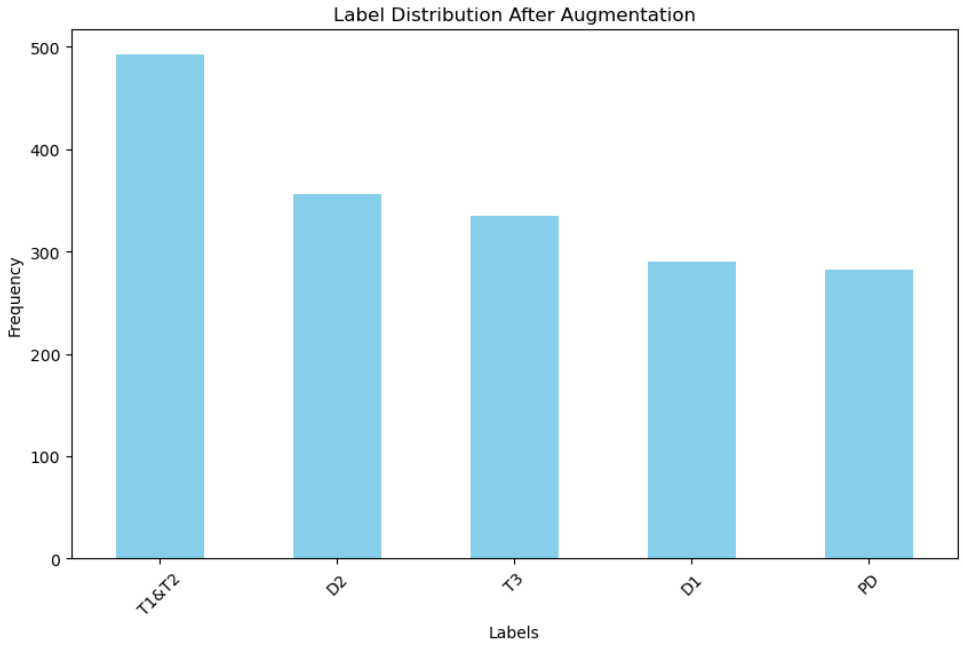

Figure 4.3: Source dataset label distribution after augmentation.

suggesting that thermal fault signatures are strongly expressed through this gas.

- **C2H6_Ratio** and **C2H2_Ratio** show substantial correlation with fault type D2 (0.63 and 0.60 respectively), indicating that high-energy discharge events are chemically distinctive.
- **H2_Ratio** exhibits moderate correlation with PD (0.48) and D1 (0.42), consistent with literature that links hydrogen emissions to electrical faults.
- **Logarithmic ratios**, such as `Ln_C2H2_C2H4` and `Ln_C2H4_C2H6`, provide additional discriminatory power by capturing nonlinear relationships – especially for D2.
- Most features show weak or negative correlation with T1&T2, highlighting the difficulty in diagnosing low and medium thermal faults using DGA alone.

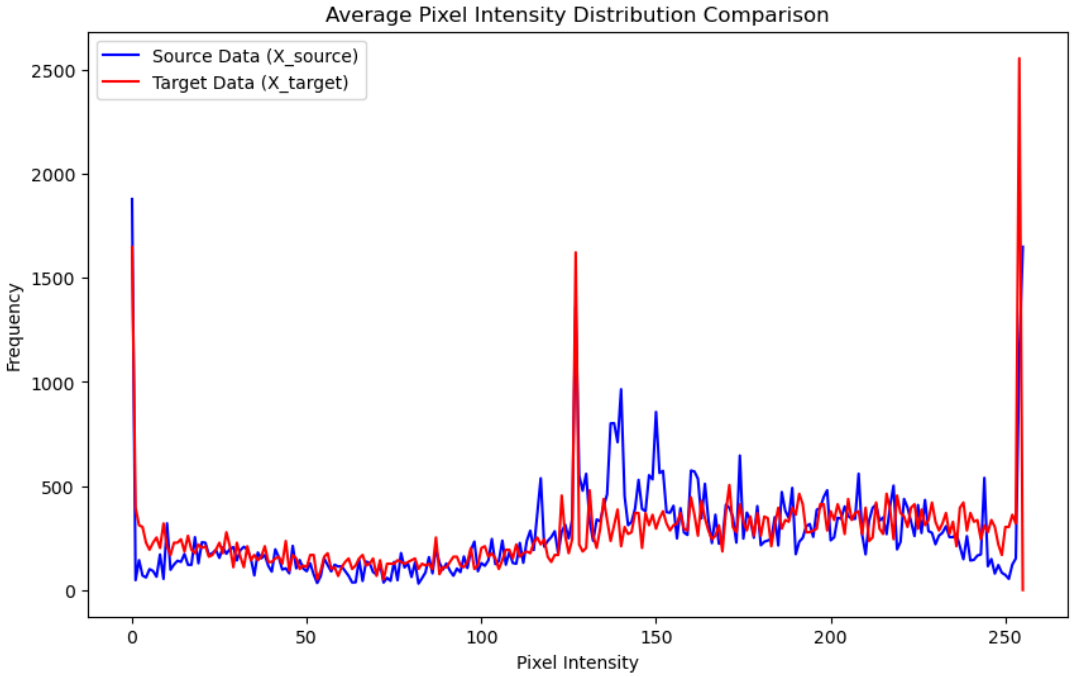

Figure 4.4: Average Pixel Intensity Distribution Comparison for GAF Images of Source and Target Datasets.

The diagonal structure in the lower portion of the matrix reflects expected identity correlations (1.00) between each fault class and itself. Off-diagonal entries in the lower-right quadrant reveal inter-class correlations – for instance, negative correlations between D2 and T1&T2 (-0.44) or PD and T3 (-0.12), suggesting mutual exclusivity in the features that define these classes.

In conclusion, this analysis confirms that the hybrid features – including both ratio and logarithmic terms – are informative and justify their inclusion in the domain adaptation process. These correlations also explain why class-wise performance varies: models have a better chance of distinguishing fault types that are more chemically distinct (e.g., T3, D2) than those with overlapping feature profiles (e.g., T1&T2).

Table 4.1: Fault Detection Performance of Different Methods

Fault Type	Accuracy (%)			F1-score (%)		
	Fine-Tuning	**MC**	**MCW**	**Fine-Tuning**	**MC**	**MCW**
D1	78.6	89.3	85.7	81.5	89.3	90.6
D2	82.1	82.1	96.4	82.1	88.5	94.7
PD	92.9	96.4	100.0	89.7	93.1	96.6
T1&T2	82.1	100.0	89.3	83.6	91.8	92.6
T3	92.9	89.3	96.4	91.2	94.3	93.1
Average	85.7	91.4	**93.6**	85.6	91.4	**93.5**

4.5 Model Performance Comparison

4.5.1 Overall Accuracy and F1-Score

The proposed MCW method significantly outperforms both Fine-Tuning and MC in terms of both accuracy and F1-score, as shown in Table 4.1. The MCW model achieves the highest average accuracy and F1-score across all five fault types.

4.5.2 Confusion Matrix Analysis

Based on Figures 4.6, 4.7 and 4.8, it can be seen that Fine-Tuning performs well on PD and T3 but suffers from significant confusion between D1 and D2. MC improves the classification of T1&T2 and reduces misclassification between D1 and D2. MCW offers the most balanced performance across all classes, further reducing confusion in T3 and achieving perfect classification in PD.

4.6 Performance under Varying Target Sample Sizes

To assess robustness under different training conditions, we evaluated model accuracy using various proportions of the target dataset. As shown in Table 4.2 and Figure 4.9, MCW consistently outperforms other methods, even when trained with as little as 30% of the target data.

Table 4.2: Accuracy (%) for Different Models Across Training Sample Sizes

Training Samples	**MCW**	**MC**	**Fine Tuning**
30%	85.9	85.0	82.1
50%	88.1	87.1	83.6
70%	93.6	91.4	85.7
90%	95.3	92.1	86.4

4.7 Hyperparameter and Architecture Sensitivity

The effect of critical hyperparameters and model configurations on classification accuracy was also examined. Our domain adaptation model is regulated by two weighting parameters, α and β, which control the relative trade-off between classification and domain alignment losses and were introduced in Equation (3.45) and Equation (3.46). To find optimal settings, several configurations were tested, as illustrated in Figure 4.10. Optimal classification accuracy was obtained when $\alpha = 0.7$ and $\beta = 0.3$, and these settings were used for further experiments.

In addition, the number of convolutional layers and fully connected (FC) layers was optimized. Figure 4.11 shows accuracy variations with different architectural configurations. The configuration with two convolutional layers and four FC layers yielded the best performance while maintaining a reasonable model complexity.

4.8 Summary of Findings

The experimental results validate the effectiveness of the proposed MCW domain adaptation framework. The key findings can be summarized as follows:

- The MCW method consistently outperforms both Fine-Tuning and MC across all fault categories in terms of accuracy and F1-score.
- Confusion matrix analysis reveals that MCW reduces inter-class misclassifications and achieves more balanced performance.

- MCW remains robust even with limited labeled target data, achieving over 85% accuracy with just 30% of target samples.
- Optimal hyperparameter settings ($\alpha = 0.7$, $\beta = 0.3$) and architectural configurations (2 CNN + 4 FC layers) were identified through thorough ablation studies.
- Distribution shift analyses (AKLD and GAF pixel histograms) confirm substantial differences between source and target domains, justifying the use of domain adaptation.

Collectively, these findings establish the MCW framework as a practical and effective solution for transformer fault diagnosis under domain shift scenarios.

4.9 Real-World Case Study: DeltaX Industrial Dataset

To validate the applicability of the proposed method in industrial settings, we incorporated a real-world case study in collaboration with DeltaX Research Inc., a Canadian company specializing in transformer diagnostics. DeltaX provided a large historical dataset containing Dissolved Gas Analysis (DGA) measurements and failure annotations across hundreds of power transformers.

The dataset exhibited substantial real-world complexity:

- Over 395,000 healthy and 6,000 failed samples, later reduced to 115,000 healthy and 1,875 failed samples after preprocessing.
- Failure labels were transformer-level and did not indicate the precise moment of failure at the row level.
- Data included metadata such as rated voltage, equipment ID, and temporal DGA readings.

Several preprocessing strategies were applied:

- Null-value filtering and outlier removal via the interquartile range method.

- Label interpretation heuristics, such as using only the final DGA row(s) to indicate transformer failure.
- Grouping by transformer-rated voltage (`ratedKv`) to reduce feature heterogeneity.
- Making an array of the final four DGA readings into a single 56-dimensional input.

Despite experimentation with classical and deep models—including Random Forest, 1D CNN, and LSTM—performance plateaued around 60–70% test accuracy, primarily due to noise, label ambiguity, and domain variance.

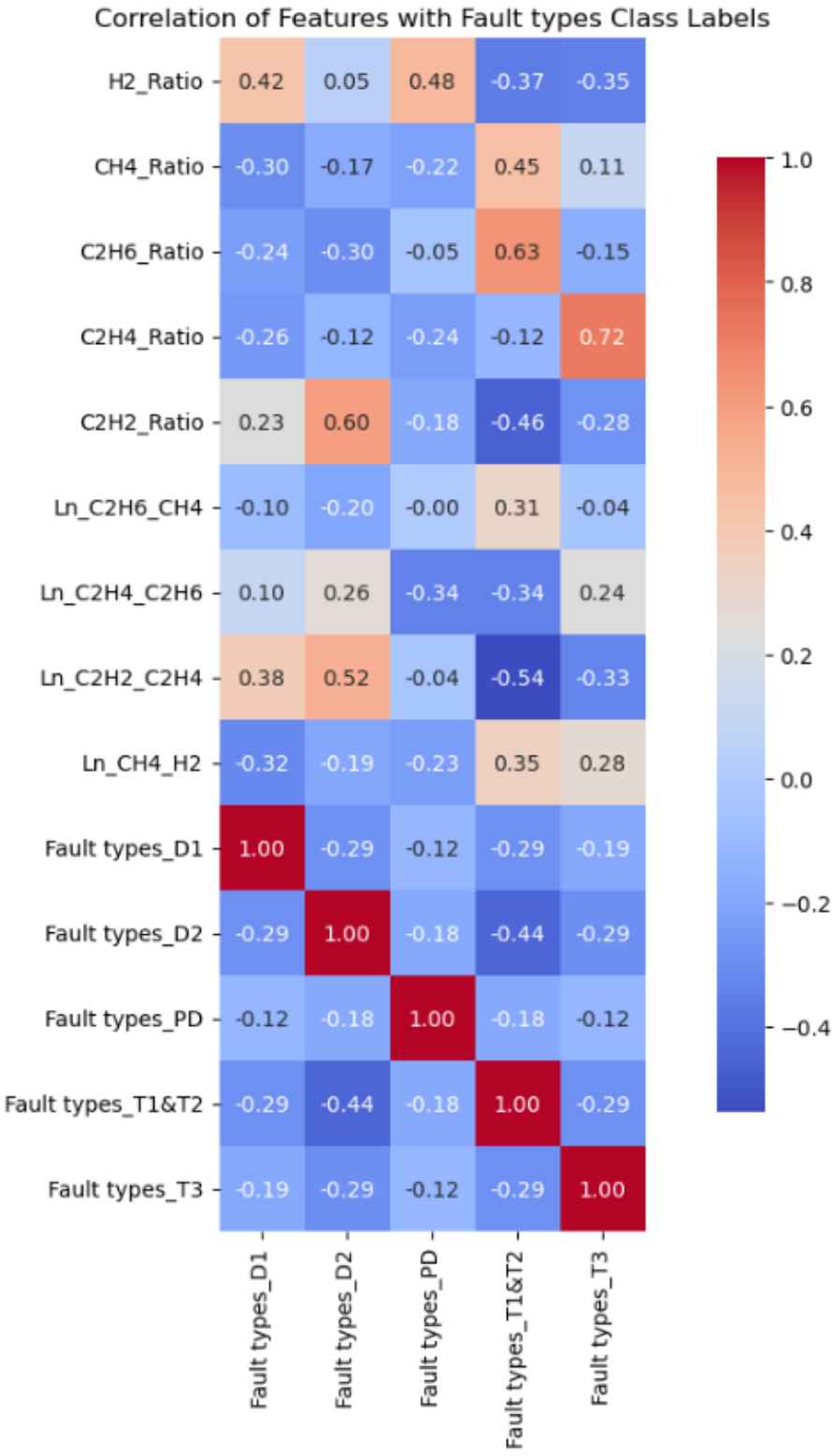

Figure 4.5: Correlation matrix between hybrid DGA features and transformer fault types. Positive correlations are shown in red, negative in blue.

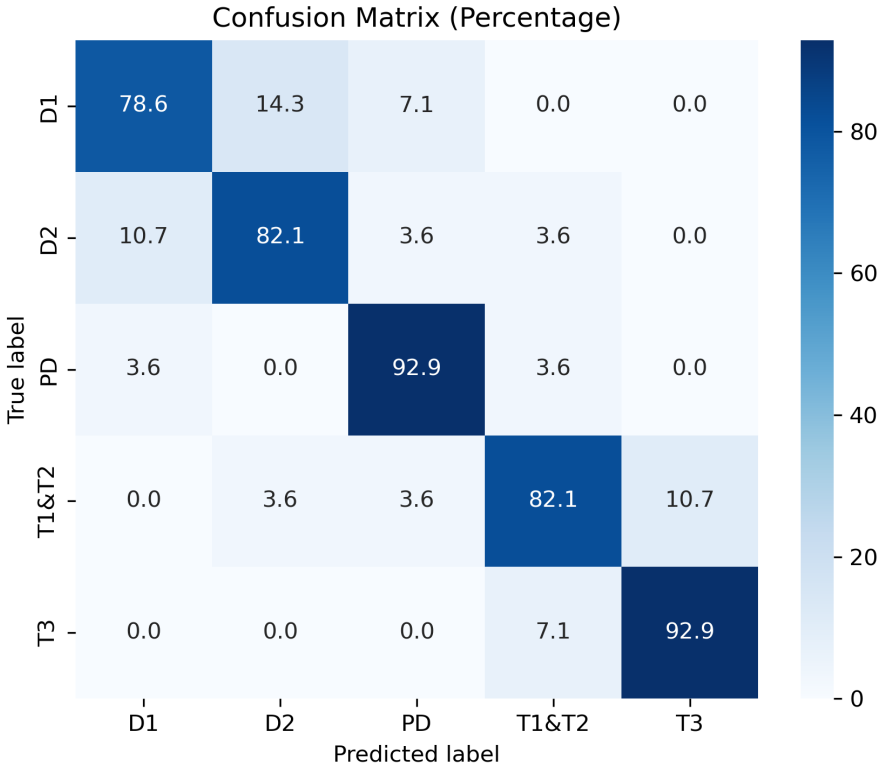

Figure 4.6: Confusion Matrix (Fine-Tuning)

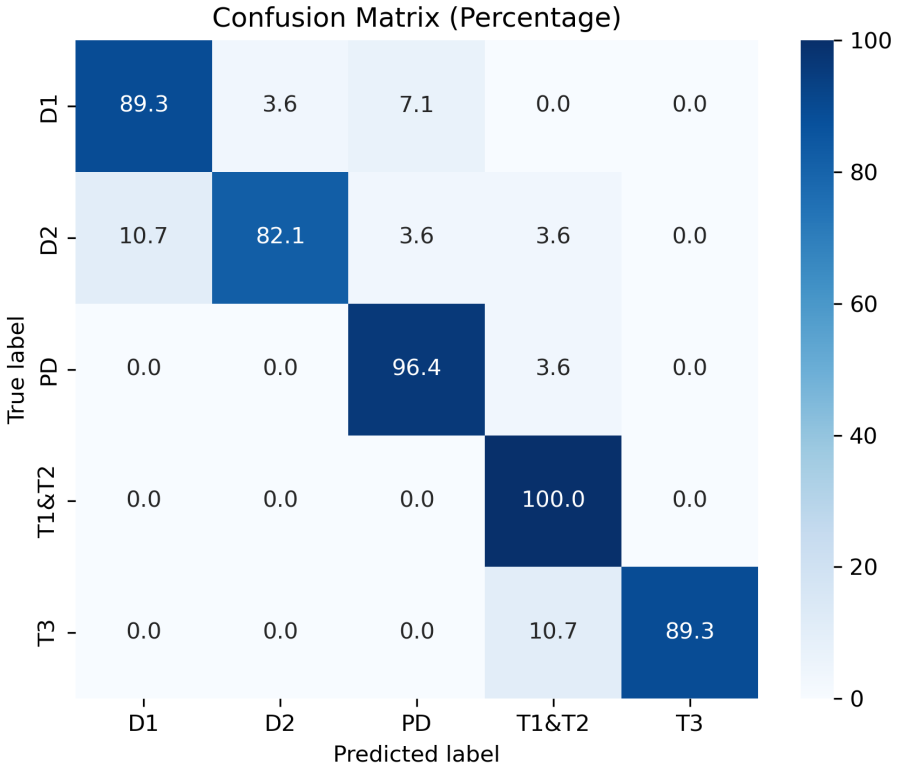

Figure 4.7: Confusion Matrix (MC)

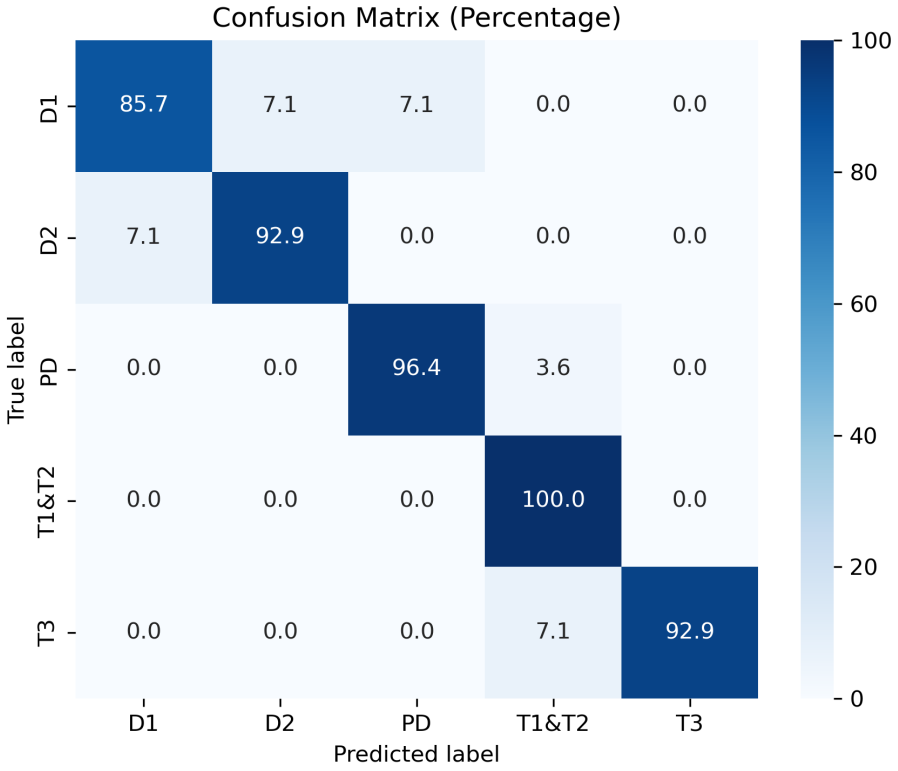

Figure 4.8: Confusion Matrix (MCW - Proposed)

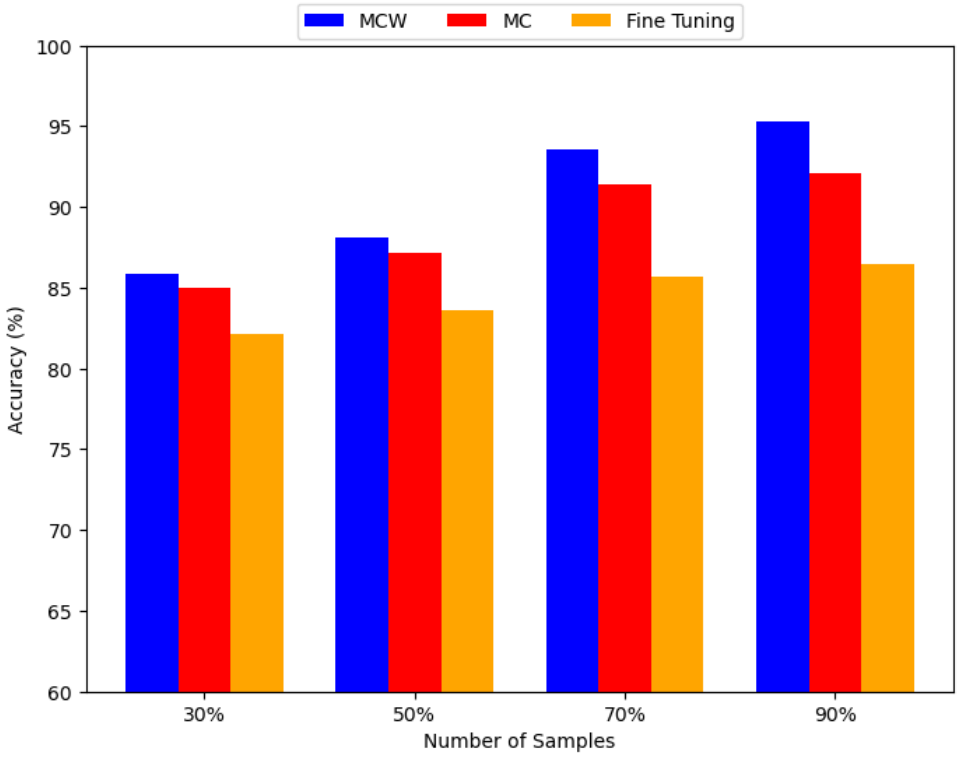

Figure 4.9: Accuracy comparison across different target sample sizes.

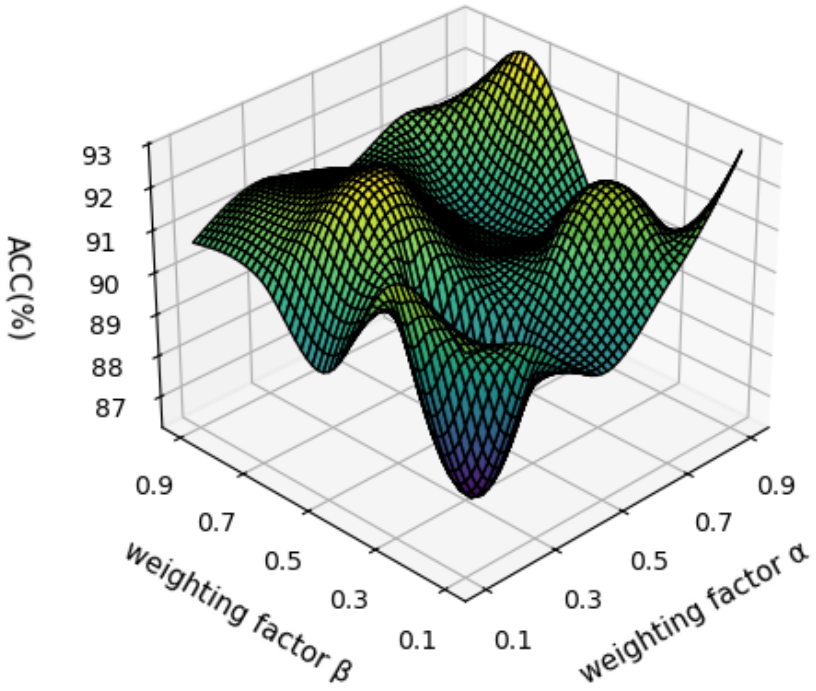

Figure 4.10: Accuracy (%) comparison for different values of α and β.

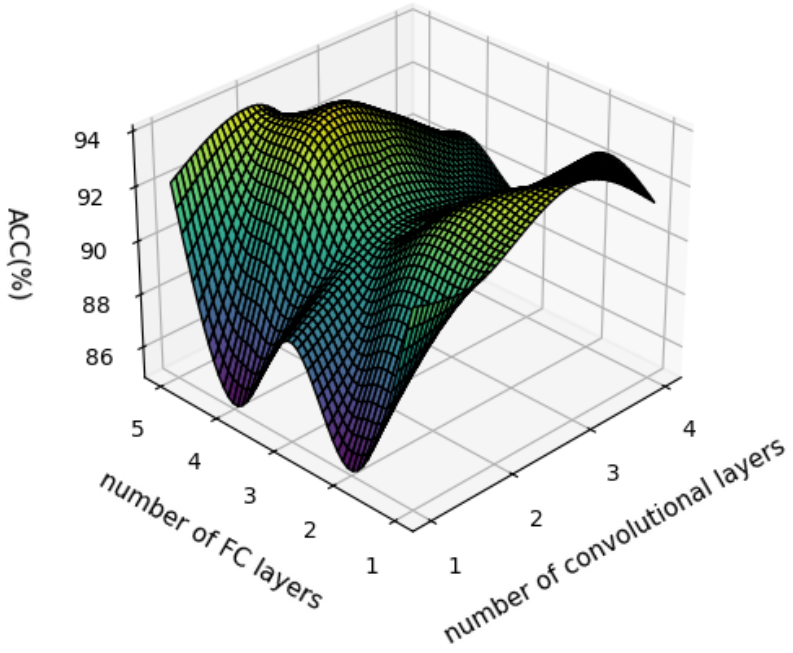

Figure 4.11: Accuracy comparison for different model architectures.

Chapter 5

Conclusion

5.1 Summary of Contributions

This thesis presented a robust and interpretable domain adaptation framework for power transformer fault diagnosis using Dissolved Gas Analysis (DGA) data. Motivated by the distribution shift between datasets collected from different utilities, the proposed method addressed a key challenge in real-world deployment: the degradation of diagnostic model performance when applied to unseen, domain-shifted data.

To overcome this challenge, we introduced a feature-weighted domain adaptation method that integrates Maximum Mean Discrepancy (MMD) and CORrelation ALignment (CORAL) losses with a novel weighting scheme derived from the Kolmogorov–Smirnov (K-S) statistic. This approach—referred to as MCW—prioritizes features exhibiting higher distributional discrepancies between the source and target domains, enabling more effective domain alignment during model training.

The methodology was validated using two distinct datasets: a source dataset compiled from Egyptian and Indian utility reports, and a target dataset derived from the IEC TC 10 database. Both datasets contained five transformer fault types: Partial Discharge (PD), Low Energy Discharge (D1), High Energy Discharge (D2), Low and Medium Thermal Fault (T1&T2), and High Thermal Fault (T3). Nine hybrid diagnostic features were extracted for each sample, combining conventional

Roger's ratios with newly derived percentage-based ratios to improve fault class separability. These features were then converted into 2D Gramian Angular Field (GAF) images, enabling visual modeling using convolutional neural networks (CNNs).

In addition to designing the proposed MCW model architecture, we conducted a comprehensive evaluation using a variety of performance metrics, including accuracy, F1-score, confusion matrices, Average Kullback–Leibler Divergence (AKLD), and pixel intensity distributions of the GAF representations. To quantify domain shift, the AKLD score of 0.698 revealed a substantial distributional gap between the source and target datasets. The pixel intensity histograms also showed that target GAF images were more uniformly distributed, whereas source GAF images were structurally sharper, justifying the necessity of domain adaptation.

5.2 Performance Insights and Comparative Evaluation

The MCW model consistently outperformed both baseline fine-tuning and the standard MC method (MMD + CORAL without feature-specific weighting) across all experiments. Confusion matrix analysis revealed that MCW yielded a more balanced and accurate classification across all five fault categories, particularly improving the classification of minority classes like D1 and D2. For instance, the MCW method achieved 100% accuracy for PD, 96.4% for D2 and T3, and a minimum per-class F1-score of 90.6%—resulting in an overall average accuracy of 93.6% and F1-score of 93.5%.

Experiments under varying target sample sizes further validated the robustness of MCW. Even with just 30% of labeled target data, MCW achieved 85.9% accuracy, outperforming both MC and fine-tuning. This demonstrates the method's effectiveness in low-data regimes, which is particularly relevant for practical deployment where labeled fault data in new domains is scarce.

5.3 Architectural Design and Ablation Studies

Extensive ablation studies were conducted to optimize key hyperparameters and architectural choices. We found that setting the domain-classification trade-off weights to $\alpha = 0.7$ and $\beta = 0.3$ led to the best performance. Furthermore, the model architecture consisting of two convolutional layers and four fully connected layers offered the optimal balance between generalization and computational efficiency. These design choices were confirmed through performance graphs presented in Chapter 4.

5.4 Scientific Implications and Future Work

The results presented in this thesis have several important scientific implications. First, they demonstrate the utility of hybrid feature engineering (combining conventional and new DGA ratios) and image-based time series transformation (GAF) in capturing discriminative patterns for fault classification. Second, the proposed MCW framework provides a generalized strategy for feature-weighted domain adaptation, offering a blueprint that could be extended to other condition monitoring problems in electrical, mechanical, or industrial systems.

However, several limitations remain. The current MCW approach relies on static K-S weights computed prior to training, which may not capture dynamic feature importance during learning. Future research could explore adaptive or learnable feature-weighting schemes within the training loop. Moreover, while CNNs proved effective in this work, integrating transformer-based architectures or attention mechanisms (e.g., Vision Transformers) may further enhance performance on complex or noisy datasets.

Another important direction lies in expanding the domain generalization capabilities of the model to unseen target domains (i.e., zero-shot settings). The inclusion of synthetic data generation techniques or self-supervised learning methods may also reduce reliance on labeled target samples.

5.5 Challenges Encountered in Real-World Deployment

Despite the encouraging results, the DeltaX case study surfaced several real-world challenges that are essential to address in future work:

- **Label Ambiguity:** The label ‘1‘ did not always correspond to the actual moment of failure, making it difficult to train supervised classifiers reliably.
- **Noise Sensitivity:** Real DGA data was noisy and susceptible to drift—small mismatches between training and testing distributions caused severe accuracy degradation.
- **Overfitting Risks:** Some feature engineering strategies, like pairwise ratios, led to artificially high training accuracy without improving generalization.
- **Input Heterogeneity:** Transformer-specific conditions (e.g., ‘ratedKv‘) introduced large variance that complicated model learning.

These limitations reaffirm the importance of robust domain adaptation strategies like MCW and open up pathways for future enhancements, including dynamic weighting, transformer architectures, or self-supervised pretraining.

5.6 Final Remarks

This research has shown that domain adaptation—when combined with interpretable, feature-weighted strategies—can significantly improve the reliability and generalizability of transformer fault diagnosis systems. By bridging the gap between source and target domains and leveraging both statistical insights and deep learning, the MCW framework contributes meaningfully to the development of intelligent, field-deployable monitoring systems for critical infrastructure.

The work laid out in this thesis paves the way for future contributions in both the academic and industrial communities focused on smart grid reliability, predictive maintenance, and fault analytics.

Bibliography

[1] Dissolved gas analysis reports. Egyptian Electr. Holding Company, Cairo, Egypt, 2016.

[2] Imran Ahmed and Xiu Lin. Temporal transformer fault analysis using cnn-lstm. *Energy AI*, 12:100263, 2023.

[3] Meng Chen and Ping Xu. A deep neural network for power transformer fault classification using dga. *IEEE Transactions on Power Delivery*, 36(2):834–841, 2021.

[4] Michel Duval. A review of faults detectable by dga and related interpretations. *IEEE Electrical Insulation Magazine*, 24(3):31–41, 2002.

[5] J. Kuffel E. Kuffel, W.S. Zaengl. *High Voltage Engineering: Fundamentals*. Newnes, 2000.

[6] Wei Fang and Min Zhang. Dga fault classification using unsupervised deep autoencoders. *Neural Computing and Applications*, 34:14261–14275, 2022.

[7] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[8] James H. Harlow. *Electric Power Transformer Engineering*. CRC Press, 2004.

[9] Jun He and Tao Li. Root cause analysis of transformer faults using xgboost and shap. *Electric Power Components and Systems*, 50(7):631–640, 2022.

[10] Martin Heathcote. *The Transformer Book: A Practical Guide to Theory and Applications*. Newnes, 1993.

[11] Wei Hong and Yujie Chen. Fault diagnosis method of power transformers based on improved variable svm. *IEEE Access*, 9:103325–103334, 2021.

[12] IEEE. Ieee std c57.104-2013: Guide for the interpretation of gases generated in oil-immersed transformers, 2013.

[13] International Electrotechnical Commission. Iec 60599: Mineral oil-impregnated electrical equipment in service – guide to the interpretation of dissolved and free gases analysis, 2015.

[14] Tao Jin and Feng Li. Transformer fault diagnosis using residual bp networks and svm-rfe. *International Journal of Electrical Power & Energy Systems*, 144:108514, 2023.

[15] Vladimir Joksimovic. *Power Transformers: Principles and Applications*. CRC Press, 2015.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[17] Anil Kumar and Samir Roy. Multi-label classification approach for overlapping transformer faults. *Expert Systems with Applications*, 217:119456, 2023.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Bo Liao and Wei Zhao. Gaf and gcn based method for dga transformer fault classification. *IEEE Access*, 8:182065–182075, 2020.

[20] X. Liu, C. Wang, Y. He, G. Yang, and Y. Hu. Fault identification for power transformer based on dissolved gas in oil data using sparse convolutional neural networks. *IET Generation, Transmission & Distribution*, 18(3):519–531, 2024.

[21] Yang Liu and Zhiqiang Hu. An ensemble classifier for transformer fault diagnosis using svm, rf, and knn. *Applied Intelligence*, 53(1):856–870, 2023.

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[23] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[24] Kensuke Mitsuzawa. Mmd-sense-analysis: Word sense detection leveraging maximum mean discrepancy. *arXiv preprint arXiv:2506.01602*, 2025.

[25] Huy Nguyen and Yu Chen. Domain adaptation for transformer fault diagnosis across utilities. *IEEE Transactions on Smart Grid*, 14(1):76–86, 2023.

[26] Farid Omar and Abhishek Singh. Transformer fault diagnosis using tabnet on dga features. *Applied Soft Computing*, 121:108752, 2022.

[27] Kiran Patel and Ankit Shah. Transformer fault detection using vision transformers and gaf images. In *2023 IEEE Conference on Smart Grid*, pages 122–127, 2023.

[28] S. Saravanan and Rakesh Gupta. Benchmarking ml classifiers for power transformer dga analysis. *IEEE Access*, 13:114325–114337, 2025.

[29] Naveen Kumar Sharma, Anuj Banshwar, Bharat Bhushan Sharma, Mohit Pathak, and Sujit Kumar. Dga-based health assessment of a 20 mva power transformer. In Kannan Govindan, Harish Kumar, and Sanjay Yadav, editors, *Advances in Mechanical and Materials Technology*, pages 779–783, Singapore, 2022. Springer Nature Singapore.

[30] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[31] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[32] Jie Sun and Xin Gao. Concurrent fault diagnosis using knowledge graph and graph neural networks. *IEEE Transactions on Industrial Informatics*, 18(4):2562–2571, 2022.

[33] Ibrahim BM Taha, Saleh Ibrahim, and Diaa-Eldin A Mansour. Power transformer fault diagnosis based on dga using a convolutional neural network with noise in measurements. *IEEE Access*, 9:111162–111170, 2021.

[34] Yiming Tang and Bin Zhao. Unsupervised transfer learning using deep coral for transformer fault classification. *IEEE Transactions on Industrial Applications*, 58(3):2832–2841, 2022.

[35] Technology Information Forecasting and Assessment Council (TIFAC).

[36] Hui Wang and Li Zhou. A one-dimensional cnn for power transformer fault classification using dga. *IET Science, Measurement & Technology*, 16(3):278–286, 2022.

[37] Yan Xiao and Tao Zhu. Attention-based transformer for dga-based transformer fault diagnosis. *Engineering Applications of Artificial Intelligence*, 120:105958, 2023.

[38] Xiaoli Yang and Lin Zhang. Intelligent diagnosis of transformer fault based on ba optimized pnn. *Electric Power Systems Research*, 171:202–210, 2019.

[39] Kai Zeng and Liang Zhou. Transformer fault diagnosis based on gwo optimized lssvm. *IET Generation, Transmission & Distribution*, 13(4):607–613, 2019.

[40] Wei Zhang, Yong Liu, Ming Chen, and Rui Gao. A fault diagnosis model of power transformers based on dissolved gas analysis features selection and improved krill herd algorithm optimized support vector machine. *Energies*, 15(7):2550, 2022.

[41] Yichen Zhang and Shan Liu. Hybrid cnn–gcn model for power transformer fault diagnosis. *Electric Power Systems Research*, 200:107989, 2021.

[42] Mei Zhao and Rui Tang. Stacked ensemble learning for transformer fault detection based on dga. *Energy Reports*, 9:411–420, 2023.

[43] Ning Zhou and Lijun Yu. Anomaly detection in transformer dga using variational autoencoders. *IEEE Transactions on Industrial Informatics*, 18(10):6954–6963, 2022.

[44] Javier Órdenes, Norman Toro, Aldo Quelopana Retamal, and A. Navarra. Data-driven dynamic simulations of gold extraction which incorporate head grade distribution statistics. *Metals*, 12:1372, 08 2022.