InvMOE: MOEs Based Invariant Representation Learning for Fault Detection in Converter Stations

Hao Sun , Shaosen Li , Hao Li , Jianxiang Huang , Zhuqiao Qiao , Jialei Wang , Xincui Tian *

Posted Date: 17 December 2024

doi: 10.20944/preprints202412.1342.v1

Keywords: converter station; fault detection; invariant learning; mixture of experts

Article

InvMOE: MOEs Based Invariant Representation Learning for Fault Detection in Converter Stations

Hao Sun [1], Shaosen Li [1], Hao Li [1], Jianxiang Huang [1], Zhuqiao Qiao [1], Jialei Wang [1], Xincui Tian [2,*]

[1] Kunming Bureau of EHV Transmission Company, Kunming, Yunnan, China
[2] Electric Power Engineering, Kunming University of Science and Technology, Kunming, Yunnan, China
* Correspondence: tianxc8866@163.com

Abstract: Converter stations are pivotal in High Voltage Direct Current (HVDC) systems, enabling power conversion between Alternating Current (AC) and Direct Current (DC), while ensuring efficient and stable energy transmission. Fault detection in converter stations is crucial for maintaining their reliability and operational safety. This paper focuses on image-based detection of five common faults: metal corrosion, discoloration of desiccant in breathers, insulator breakage, hanging foreign objects, and valve cooling water leakage. Despite advancements in deep learning, existing detection methods face two major challenges: limited model generalization due to diverse and complex backgrounds in converter station environments, and sparse supervision signals caused by the high cost of collecting labeled images for certain faults. To overcome these issues, we propose InvMOE, a novel fault detection algorithm with two core components: (1) invariant representation learning, which captures task-relevant features and mitigates background noise interference, and (2) multi-task training using a mixture of experts (MOE) framework to adaptively optimize feature learning across tasks and address label sparsity. Experimental results on real-world datasets demonstrate that InvMOE achieves superior generalization performance and significantly improves detection accuracy for tasks with limited samples, such as valve cooling water leakage. This work provides a robust and scalable approach for enhancing fault detection in converter stations.

Keywords: converter station; fault detection; invariant learning; mixture of experts

1. Introduction

Converter stations play a pivotal role in High Voltage Direct Current (HVDC) systems by enabling efficient power conversion between Alternating Current (AC) and Direct Current (DC). They regulate power flow, ensuring stable and reliable energy transmission across interconnected grids, which is essential for long-distance power transmission, asynchronous grid interconnection, and cross-regional power distribution. The critical importance of these systems cannot be overstated, as failures can lead to widespread power outages, grid instability, and significant economic disruptions [1,2]. Consequently, effective fault detection is indispensable to maintain the reliable operation of converter stations and the overall health of the power network.

However, fault detection in converter stations is inherently challenging due to the complex and dynamic environments in which these stations operate. The environments are subject to various unpredictable factors such as changing lighting conditions, weather, camera angles, and background clutter, which can drastically affect the performance of conventional detection models. Traditional fault detection methods, which often rely on domain-specific expertise and handcrafted features [3,4], are limited in their ability to adapt to such environmental variability. These methods typically involve feature extraction based on fixed rules, which do not generalize well under different conditions, leading to performance degradation when deployed in real-world scenarios.

In contrast, deep learning-based methods, particularly those utilizing large-scale data, have shown significant promise in automatically learning rich, high-level feature representations [5,6]. Convolutional Neural Networks (CNNs) have been widely used for tasks such as feature extraction, fusion, and decision-making. While these methods have demonstrated impressive performance on controlled datasets, they are not immune to two critical limitations that hinder their real-world applicability: **(1) Limited Model Generalization**: Environmental variations, such as changes in lighting,

 doi:10.20944/preprints202412.1342.v1

weather, and the introduction of background noise, can significantly degrade model performance. Models trained on specific datasets may fail to generalize to out-of-distribution (OOD) data, leading to poor performance when the operational conditions differ from those encountered during training [7,8]. This challenge is particularly pronounced in converter station environments, where conditions can vary unpredictably over time. **(2) Sparse Supervision Signals**: Fault detection tasks are heavily reliant on labeled data, but acquiring high-quality annotations for certain fault categories, such as valve cooling water leakage, is expensive and time-consuming. As a result, there is often a severe imbalance in the availability of labeled data, which makes it difficult to train robust models [9,10]. Sparse supervision exacerbates the difficulty of accurately detecting rare or hard-to-label faults.

To address these challenges, we introduce **InvMOE** (**Inv**ariant representation learning with **M**ixture **O**f **E**xperts), a novel fault detection framework specifically designed for the dynamic and challenging environments of converter stations. InvMOE incorporates two key components: **(1) Invariant Representation Learning**: This component aims to disentangle task-relevant features from environmental noise, thereby enhancing the model's robustness and generalization capabilities. From a causal perspective, fault occurrences are independent of environmental factors, which are treated as confounders [11]. By learning invariant representations, the model can focus on the causal features that are consistent across different environments, in alignment with causal inference principles. Techniques such as adversarial training and contrastive loss are employed to facilitate this disentanglement, making InvMOE particularly well-suited for OOD scenarios where environmental factors vary significantly from the training data distribution. **(2) Multi-task Training with Mixture of Experts (MOE)**: To tackle the issue of sparse supervision, the MOE framework is employed to adaptively route inputs to specialized expert subnetworks that are optimized for different fault detection tasks. This approach not only prevents negative transfer between tasks but also allows the model to leverage shared knowledge across tasks, improving performance even for categories with limited data. By simultaneously learning from multiple related tasks, the model can effectively utilize available data, mitigating the challenge of data scarcity for rare fault types, such as valve cooling water leakage detection.

Experiments conducted on real-world datasets demonstrate that InvMOE significantly outperforms existing methods, achieving superior generalization across OOD conditions and showing substantial improvements in fault detection tasks with limited labeled data.

Contributions: The main contributions of this paper are as follows:

- We propose InvMOE, a novel fault detection algorithm that combines invariant representation learning and a mixture of experts framework to address the challenges of limited generalization and sparse supervision.
- We introduce a causal-inspired approach for disentangling task-relevant features from environmental noise, enabling robust and reliable fault detection across diverse and unpredictable converter station environments.
- We develop a multi-task training strategy with MOE, which improves model efficiency and effectiveness, particularly in handling tasks with limited data, and demonstrate its superiority through extensive experiments on real-world datasets.

2. Related Works

In this section, we review the related works on fault detection methods for converter stations, out-of-distribution (OOD) generalization techniques, and multi-task learning approaches, with a specific focus on image recognition and detection tasks.

2.1. Converter Station Fault Detection

Fault detection in converter stations is a critical task to ensure the safe and efficient operation of high-voltage direct current (HVDC) systems [1]. Traditional fault detection methods for converter stations often rely on rule-based systems, threshold-based approaches, or model-based techniques [3,4]. These methods typically focus on detecting specific anomalies like overcurrent or voltage irregularities,

 doi:10.20944/preprints202412.1342.v1

which might indicate faults in key components such as transformers, capacitors, or switches. For instance, methods like open circuit fault diagnosis have been used for detecting anomalies in converter circuits, with some systems employing automatic feature extraction coupled with algorithms such as random forests to identify faults in the presence of nonstationary influences [3]. More recent approaches have leveraged deep learning [5] and computer vision techniques [6] for fault detection in converter stations, particularly focusing on analyzing images captured in real-world environments [12]. These methods generally involve extracting features from images and classifying them into fault categories. Additionally, researchers have explored hybrid techniques that combine deep learning with traditional electrical fault detection to improve diagnostic accuracy and robustness [13]. Furthermore, some studies have extended fault detection by incorporating environmental factors, such as light conditions and camera angles, which can significantly affect the accuracy of image-based fault detection systems. This incorporation of contextual features helps improve the robustness and generalization of fault detection models, addressing the challenges of real-world deployment [14].

2.2. Out-of-Distribution Generalization

Out-of-Distribution (OOD) generalization is a significant challenge in machine learning [15,16], especially when deploying models in dynamic real-world environments like converter stations, where variations in lighting, angle, and background can drastically impact model performance. OOD generalization refers to the model's ability to make accurate predictions on data that differs from the training data distribution. Several approaches have been proposed to address this challenge, with a strong focus on domain adaptation and robust learning techniques. Invariant Risk Minimization (IRM) is one of the prominent techniques used to address OOD generalization [8,17–19]. IRM encourages the model to learn invariant features across different domains (or environments) to ensure consistency in predictions when exposed to unseen data distributions. This is particularly useful when labeled data from the target domain (e.g., converter stations under specific environmental conditions) is scarce. Techniques like adversarial training and self-supervised learning have been utilized to align feature distributions between source and target domains, reducing the domain shift that hampers generalization [20,21]. Additionally, methods like data augmentation (e.g., random cropping, rotation, color jittering) are commonly used to expose models to a variety of input conditions during training. This technique helps improve the model's robustness, allowing it to better handle OOD data when deployed in diverse operational environments.

2.3. Multi-task Learning

Multi-task learning (MTL) is a machine learning paradigm that simultaneously learns multiple tasks with shared representations, aiming to improve the model's performance on each individual task by leveraging common information [22]. MTL has been widely applied to image recognition, classification, and detection tasks, where multiple related objectives need to be tackled simultaneously.

In the context of fault detection in converter stations, MTL can be particularly beneficial, as it allows the model to learn different fault detection tasks (e.g., corrosion detection, insulator breakage detection) concurrently, sharing useful features across tasks. For instance, one task might focus on detecting corrosion, while another focuses on detecting insulator cracks. By sharing the learned features, the model benefits from a more generalized understanding of the environment, which can improve performance on each individual task, especially when the amount of labeled data is limited for some tasks [20] is a popular approach within MTL, where different "experts" specialize in different tasks and are dynamically selected based on the input. MOE provides the flexibility of task specialization while maintaining the benefits of sharing common knowledge. In image recognition tasks, the MOE framework has been applied to ensure that the relevant experts are activated depending on the type of fault being detected. This allows the model to efficiently allocate resources to tasks that require more specialized attention, while still benefiting from shared feature learning for tasks that have overlapping characteristics. MTL methods can be combined with techniques like attention

 doi:10.20944/preprints202412.1342.v1

mechanisms and graph neural networks (GNNs) to model dependencies between different faults or tasks in converter station environments [21]. These dependencies can enhance the model's ability to handle complex relationships between fault types, improving detection accuracy.

3. The Proposed InvMOE Framework

This section describes the proposed **InvMOE** framework for fault detection in converter stations. The method consists of three key components: (1) image feature extraction, (2) MOE-based multi-task learning, and (3) invariant learning based optimization. The detailed processes of each component are as follows.

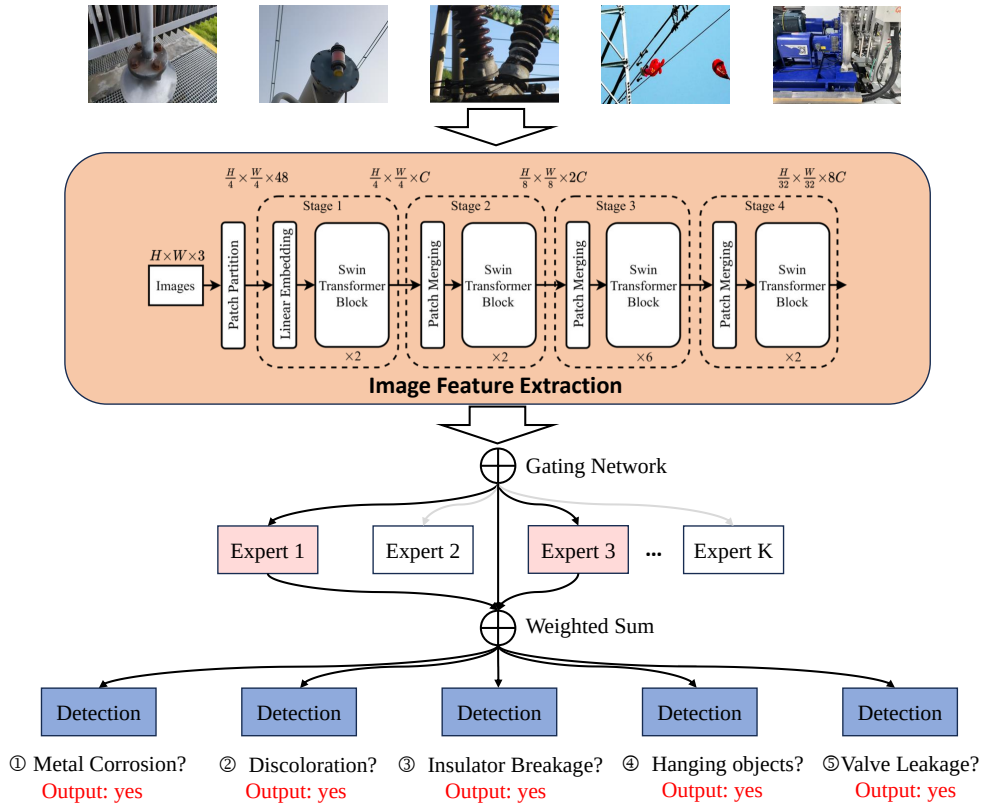

Figure 1. Overview of our proposed InvMOE framework.

3.1. Image Feature Extraction

The first stage of the **InvMOE** framework focuses on extracting robust feature representations from images captured in real-world converter station environments. These images, taken under varying conditions of lighting, angles, and background complexity, serve as the primary data source for fault detection.

While the framework is designed to accommodate any state-of-the-art visual backbone, we employ the **Swin Transformer** model due to its advanced capabilities in capturing both local and global features through a hierarchical structure and self-attention mechanism [23]. The Swin Transformer has shown superior performance in a variety of vision tasks, making it an ideal choice for the diverse and challenging nature of converter station images [10,23].

Given an input image $x \in \mathbb{R}^{H\times W\times C}$, representing a high-resolution RGB photo with height H, width W, and C color channels, the Swin Transformer processes the image as follows:

- **Patch Tokenization:** The input image x is divided into non-overlapping patches of size $P \times P$. Each patch is flattened into a vector, and a linear embedding layer maps these vectors into a d-dimensional feature space. This produces an initial set of tokens $\mathbf{T} \in \mathbb{R}^{(H/P)\times(W/P)\times d}$.

 doi:10.20944/preprints202412.1342.v1

- **Hierarchical Feature Extraction:** The tokenized patches are passed through multiple transformer layers. Each layer consists of *shifted window-based self-attention* modules and feed-forward networks, enabling efficient computation and the capture of long-range dependencies within the image [23].
- **Feature Aggregation:** As the processing progresses through hierarchical stages, features are aggregated and down-sampled to reduce spatial dimensions while increasing semantic richness. This results in a compact feature vector $\mathbf{z} \in \mathbb{R}^d$, encapsulating the key visual information of the input image.

The extracted feature vector $\mathbf{z}$ serves as the input for subsequent stages of the **InvMOE** framework, ensuring that rich and invariant representations are available for fault detection tasks. By leveraging the Swin Transformer's ability to effectively balance computational efficiency and representation quality, this stage establishes a strong foundation for the proposed method.

3.2. MOE-Based Multi-Task Learning

InvMOE leverages a **mixture of experts (MOE)** framework to address the challenges posed by multiple fault detection tasks and sparse supervision signals. For each input image feature, extracted via a Swin Transformer, the MOE architecture is used to adapt the features for the downstream five fault detection tasks: metal corrosion, respirator silicone discoloration, insulator fracture, suspended object, and valve cooling water leakage detection.

3.2.1. Adaptive Expert Routing

The generalized features $\mathbf{z}_c$ obtained from the invariant representation learning stage are passed through the MOE layer. The MOE framework consists of K expert networks $\{E_1, E_2, \ldots, E_K\}$, where each expert specializes in a subset of fault detection tasks. A gating network G dynamically routes the input to one or more experts based on the task requirements:

$$\mathbf{z}_{\text{output}} = \sum_{i=1}^{k} g_i(\mathbf{z}) E_i(\mathbf{z}), \tag{1}$$

where $g_i(\mathbf{z})$ is the gating weight for expert E_i. This adaptive routing mechanism enables the model to allocate computational resources effectively, ensuring that each task benefits from task-specific expertise. In this case, the five tasks will use a combination of experts to process their respective features, ensuring specialized detection.

3.2.2. Multi-task learning

To jointly train the experts for all fault detection tasks, we adopt an empirical risk minimization (ERM) based multi-task optimization framework. The loss function for each individual task, $i \in \{1, \ldots, 5\}$, is based on the cross-entropy loss, defined as:

$$\mathcal{L}_{\text{task}_i} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{2}$$

where y_c is the ground truth for class c for task i, p_c is the predicted probability for class c, and C is the number of possible classes for the task. The model has five such cross-entropy loss functions, one for each task.

The overall multi-task loss function is a weighted sum of the task-specific losses:

$$\mathcal{L}_{\text{ERM}} = \sum_{i=1}^{5} \lambda_i \mathcal{L}_{\text{task}_i}, \tag{3}$$

 doi:10.20944/preprints202412.1342.v1

where λ_i is the weight for task i, $\mathcal{L}_{\text{ERM}}$ denote the total empirical risk minimization. This design not only enhances the performance of individual tasks but also facilitates knowledge sharing across tasks, leveraging common features to mitigate the issue of sparse supervision. The weighted sum of losses ensures the model optimizes all five tasks simultaneously, with task-specific contributions adjusted according to the importance and difficulty of each task.

3.3. Invariant Learning-based Optimization

To address the challenge of limited generalization caused by environmental variability, we adopt an **invariant representation learning** strategy inspired by Invariant Risk Minimization (IRM). Specifically, we divide the input images into multiple environments $\{e_1, e_2, \ldots, e_N\}$ based on factors such as lighting, angles, and surrounding backgrounds. The goal is to encourage the model to focus on causal features that are invariant across these environments while ignoring task-irrelevant environmental noise.

Algorithm 1 InvMOE Framework for Fault Detection

Input: Image $x \in \mathbb{R}^{H \times W \times C}$
Output: Fault detection results for five tasks
Step 1: Feature Extraction
  Divide x into patches and apply linear embedding to map to d-dimensional space.
  Pass through Swin Transformer and aggregate features to get $\mathbf{z} \in \mathbb{R}^d$.
Step 2: MOE-based Multi-task Learning
  Pass $\mathbf{z}$ through MOE layer with K experts, using dynamic routing by gating network.
  For each task i, compute cross-entropy loss $\mathcal{L}_{\text{task}_i}$ and total loss $\mathcal{L}_{\text{ERM}}$.
Step 3: Invariant Learning-based Optimization
  Model $\mathbf{z}$ as causal and environmental factors, extract causal factors $\mathbf{z}_c$.
  Minimize variance across environments with IRM regularization.
Step 4: Parameter Optimization
  Optimize model parameters using gradient-based optimization (e.g., Adam optimizer) to minimize the total loss:
  $\mathcal{L} = \mathcal{L}_{\text{ERM}} + \beta \text{Var}(\{\mathcal{L}^k_{\text{ERM}}\})$
Step 5: Output
  Output fault detection results for five tasks.

3.3.1. Causal Framework for Invariance

From a causal perspective, the fault label y is determined by latent causal factors $\mathbf{z}_c$, independent of environmental factors $\mathbf{z}_e$. We model the representation $\mathbf{z}$ as the combination of $\mathbf{z}_c$ and $\mathbf{z}_e$, where:

$$\mathbf{z} = \mathbf{z}_c + \mathbf{z}_e, \tag{4}$$

The goal of invariant representation learning is to extract $\mathbf{z}_c$ while suppressing $\mathbf{z}_e$, thus focusing on the causal factors that are responsible for fault detection and minimizing the influence of environmental variations.

3.3.2. IRM-Based Regularization

We employ Invariant Risk Minimization (IRM) to enforce invariance across different environments. The IRM framework aims to encourage the model to learn the stable detection ability across various environments. Thus, the IRM-based regularization is defined as follows:

$$\mathcal{L}_{\text{IRM}} = \text{Var}(\{\mathcal{L}^k_{\text{ERM}} : 1 \leq k \leq K\}), \tag{5}$$

where $\mathcal{L}^k_{\text{IRM}}$ is the loss for environment e_k. This formulation encourages the model to minimize the loss across all environments, ensuring that the learned representations are invariant and robust to environmental variations.

By combining the IRM-based regularization, the final optimization objective of **InvMOE** is:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}^k_{\text{ERM}} + \beta \text{Var}(\{\mathcal{L}^k_{\text{ERM}} : 1 \leq k \leq K\}), \tag{6}$$

 doi:10.20944/preprints202412.1342.v1

where β is the hyper-parameter to balance the ERM and IRM losses. Based on the above optimization, **InvMOE** can learn the invariant representation $\mathbf{z}_c$, and achieve robust fault detection performance in converter station environments, addressing both the issues of sparse supervision and limited model generalization due to environmental variability. The detailed training process as shown in Algorithm 1.

4. Experimental Results

4.1. Dataset Preprocessing

In order to evaluate the performance of the proposed **InvMOE** framework for fault detection in converter stations, we utilize a custom dataset consisting of high-resolution RGB images captured from various converter stations under real-world conditions. The dataset includes images of different fault types, such as metal corrosion, respirator silicone discoloration, insulator fracture, suspended objects, and valve cooling water leakage. Given the challenging nature of these images, the dataset processing steps are carefully designed to ensure high-quality and diverse inputs for training the model.

4.1.1. Data Collection

The dataset is collected from multiple converter stations across varying environmental conditions, including different lighting, camera angles, and background complexities. Each image is annotated with labels corresponding to the five fault detection tasks. The dataset is split into training, validation, and test sets, with approximately 80% of the data allocated for training, 10% for validation, and 10% for testing. The detailed distribution of images in the dataset is summarized in Table 1.

4.1.2. Data Augmentation

To enhance the generalization capabilities of the model and mitigate the effects of sparse supervision, data augmentation techniques are applied to increase the diversity of the training data. These techniques include:

- **Random cropping:** Randomly cropping regions of the image to simulate variations in object size and position.
- **Color jittering:** Random adjustments to the image's brightness, contrast, and saturation to simulate varying lighting conditions.
- **Rotation and flipping:** Random rotations and horizontal flips to simulate different camera angles.
- **Noise injection:** Adding random noise to images to simulate real-world disturbances and background complexities.

4.1.3. Normalization and Standardization

Before feeding the images into the model, the pixel values are normalized to the range $[0, 1]$ by dividing by 255. Additionally, the images are standardized by subtracting the mean and dividing by the standard deviation of the training dataset to ensure consistent feature scaling. This helps to improve the convergence of the model during training and ensures that the input data is suitable for the Swin Transformer.

4.2. Experimental Settings

4.2.1. Training Setup

To ensure stable and efficient training, the following setup is used:

- The model is trained using mini-batch gradient descent with a batch size of 32.
- Early stopping is employed to prevent overfitting, with a patience of 10 epochs. Training stops if the validation loss does not improve for 10 consecutive epochs.
- The learning rate is initialized at 0.001 and decreased by a factor of 0.1 after every 20 epochs.

 doi:10.20944/preprints202412.1342.v1

- The Adam optimizer is used for model optimization, which adapts the learning rate based on first and second moments of the gradients.

Table 1. Dataset Statistics: Pu'er Converter Station Dataset.

Task	Number of Images
Task 1: Metal Corrosion	500
Task 2: Silica Gel Discoloration	500
Task 3: Insulator Breakage	500
Task 4: Overhead Suspension	500
Task 5: Valve Cooling Water Leak	100

4.2.2. Evaluation Baselines

We conduct experiments with several competing baselines, which are introduced as follows:

- **ResNet:** ResNet (Residual Networks) [6] is a deep convolutional neural network architecture known for its use of residual connections, which help mitigate the vanishing gradient problem by allowing gradients to flow through the network more effectively. It is particularly effective in image classification tasks and has been widely used in various computer vision applications. ResNet is commonly employed as a baseline model for comparison in tasks requiring deep learning architectures.
- **Swin Transformer:** The Swin Transformer [23] is a state-of-the-art vision transformer architecture that uses shifted windows for efficient self-attention and hierarchical feature representation. It overcomes the limitations of traditional Vision Transformers (ViTs) by processing images in smaller, non-overlapping patches and dynamically adjusting attention regions, making it highly effective for capturing both local and global features. It has shown superior performance in various vision tasks compared to CNN-based architectures.
- **IRM (Invariant Risk Minimization):** IRM [8] focuses on learning representations that generalize across multiple environments by enforcing invariance in the learned features. In the context of fault detection, this method would remove the multi-task learning component, resulting in a model variant that learns invariant representations across different environmental conditions without task-specific adaptation. This approach helps in addressing environmental variability, but without the benefit of multi-task learning shared across tasks.

4.2.3. Evaluation Metrics

The performance of the model is evaluated using two key metrics: accuracy (ACC) and F1-score. These metrics are defined as follows:

- **Accuracy (ACC):** The accuracy of a model is the proportion of correct predictions out of the total number of predictions. It is computed as:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

  where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively. Accuracy provides a general measure of how well the model is performing across all classes.
- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

 doi:10.20944/preprints202412.1342.v1

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The F1-score is especially useful when the dataset is imbalanced, as it accounts for both false positives and false negatives.

4.3. Performance Comparisons with Baselines

We evaluate the performance of the proposed **InvMOE** model on five distinct fault detection tasks, covering scenarios such as detecting metal corrosion, respirator silica gel discoloration, insulator breakage, overhead suspension detection, and valve cooling water leakage.

Table 2. Accuracy of different models for fault detection tasks (%).

Model	Task 1	Task 2	Task 3	Task 4	Task 5
ResNet	94.0	93.2	91.5	92.0	80.0
Swin Transformer	96.5	95.3	94.1	94.5	84.2
IRM	97.0	96.5	95.0	94.8	85.5
InvMOE	**97.5**	**96.8**	**95.6**	**95.1**	**88.0**

Table 3. F1-Score of different models for fault detection tasks (%).

Model	Task 1	Task 2	Task 3	Task 4	Task 5
ResNet	93.5	92.0	90.3	90.8	75.2
Swin Transformer	96.2	94.4	93.0	93.2	80.1
IRM	97.1	96.4	94.9	94.7	82.5
InvMOE	**97.3**	**96.5**	**95.4**	**94.9**	**87.5**

The performance of the proposed **InvMOE** model was compared to three baseline models: ResNet, Swin Transformer, and IRM, across five distinct fault detection tasks. The accuracy and F1-score results, as shown in Tables 2 and 3, reveal several key findings:

- **Accuracy Performance**: InvMOE consistently outperforms all baseline models across all tasks, achieving the highest accuracy in each task. Notably, InvMOE achieves an accuracy of 97.5% in Task 1 (metal corrosion detection), which is higher than the next best model, IRM (97.0%), and significantly higher than ResNet (94.0%) and Swin Transformer (96.5%). In Task 5 (valve cooling water leakage), InvMOE maintains an impressive accuracy of 88.0%, surpassing all baseline models, with IRM coming in second at 85.5%.
- **F1-Score Performance**: The trend observed in the accuracy results is reflected in the F1-scores. InvMOE again leads with the highest F1 scores across all tasks. For example, in Task 1, InvMOE achieves an F1-score of 97.3%, outperforming IRM (97.1%), Swin Transformer (96.2%), and ResNet (93.5%). In Task 5, InvMOE maintains its superior performance with an F1-score of 87.5%, which is considerably higher than IRM (82.5%) and Swin Transformer (80.1%).
- **Comparison to Baselines**: ResNet, while a strong baseline, generally falls behind both Swin Transformer and IRM in terms of both accuracy and F1-score. This is expected given that ResNet is a convolutional neural network, which may not capture the fine-grained relationships and long-range dependencies in the data as effectively as transformer-based models. Swin Transformer and IRM perform similarly on most tasks, with IRM slightly outperforming Swin Transformer. This indicates that enforcing invariance across different environments (as done by IRM) offers some benefits over the self-attention mechanism used in Swin Transformer, particularly in tasks with varied environmental conditions. Overall, InvMOE demonstrates the most robust performance, suggesting that the integration of multi-task learning and the model's ability to handle diverse fault detection scenarios contribute to its superior results.

 doi:10.20944/preprints202412.1342.v1

In summary, **InvMOE** significantly outperforms all baseline models in both accuracy and F1-score across the fault detection tasks, confirming its effectiveness in real-world fault detection applications.

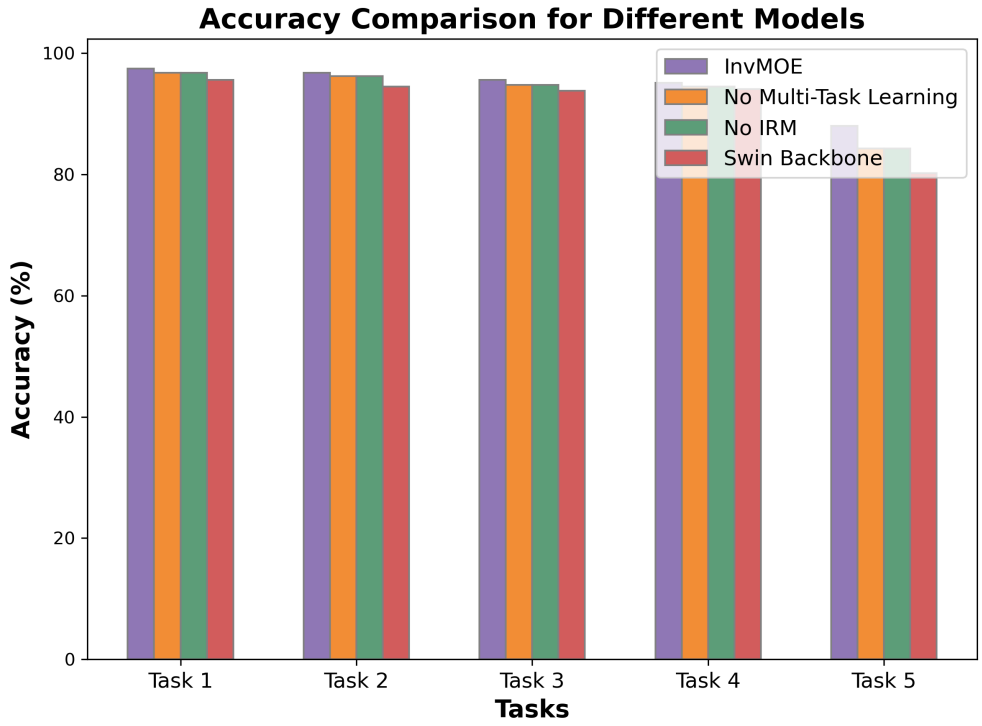

Figure 2. Accuracy Comparison for Different Variants.

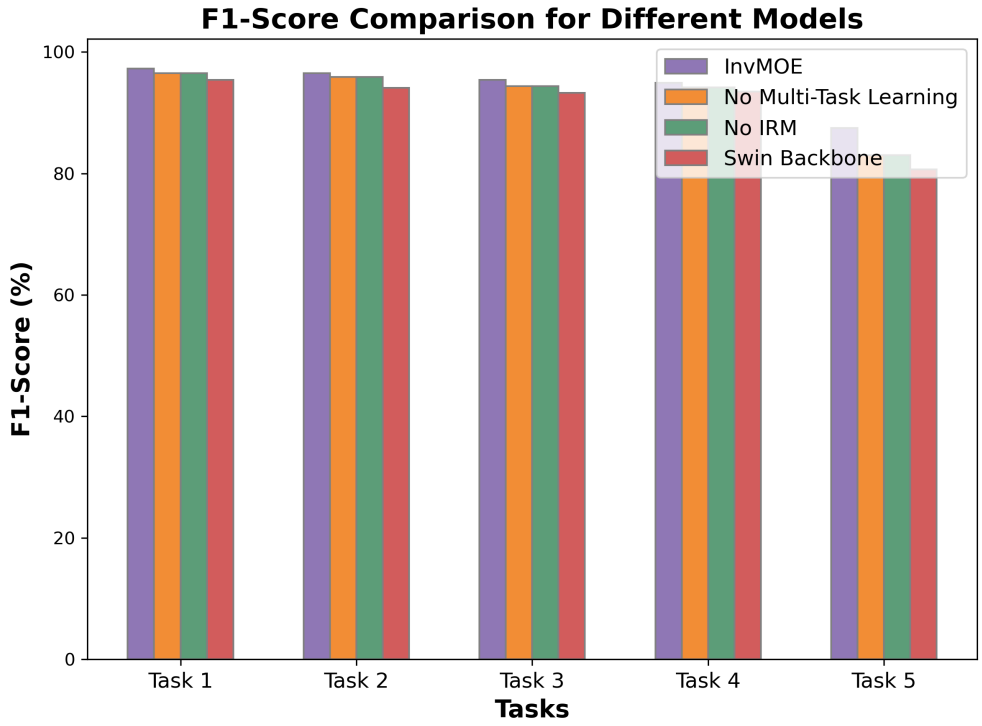

Figure 3. F1-Score Comparison for Different Variants.

 doi:10.20944/preprints202412.1342.v1

5. Ablation Study

In this section, we conduct an ablation study to evaluate the contributions of key components in the **InvMOE** model. Specifically, we assess the effect of removing multi-task learning, the Invariant Risk Minimization (IRM) mechanism, and the Swin Transformer backbone on the model's performance. The following variants are considered:

- **InvMOE (full model)**: The complete model that incorporates multi-task learning, IRM, and the Swin Transformer backbone.
- **No Multi-Task Learning**: In this variant, the multi-task learning component is removed, and the model is trained in a single-task learning setup.
- **No IRM**: This variant removes the IRM mechanism, leaving the model to train without enforcing invariant risk minimization.
- **Swin Transformer Backbone**: In this variant, we keep the multi-task learning and IRM components but replace the Swin Transformer backbone with a simpler architecture for comparison.

5.1. Ablation Study Results

The accuracy and F1-score results for the different variants of the **InvMOE** model are summarized in Figure 2 and Figure 3. We can observe the followings:

- **Impact of Multi-Task Learning**: When multi-task learning is removed, the model's accuracy and F1-score decrease by around 1-2% for all tasks. This demonstrates the importance of leveraging shared information from multiple tasks to improve generalization and performance across different fault detection scenarios.
- **Impact of Invariant Risk Minimization (IRM)**: Similarly, removing the IRM component causes a noticeable reduction in accuracy and F1-score (around 1-2%), which suggests that the IRM mechanism is crucial for mitigating the effects of variability across different environments. Its absence leads to slightly less stable performance on some tasks.
- **Impact of Swin Transformer Backbone**: Replacing the Swin Transformer backbone with a simpler architecture causes the largest performance drop, particularly in the more complex tasks. The decrease in accuracy and F1-score highlights the strength of the self-attention mechanism and hierarchical feature extraction of the Swin Transformer, which allows the model to capture long-range dependencies and contextual information more effectively than simpler architectures.

In conclusion, the ablation study confirms that the full **InvMOE** model, with its combination of multi-task learning, IRM, and the Swin Transformer backbone, provides the best performance across all tasks. Removing any of these components leads to a reduction in model performance, underscoring the importance of each element in the overall effectiveness of the model.

6. Conclusion

In this study, we proposed **InvMOE**, an advanced fault detection algorithm designed to tackle the challenges of generalization and sparse supervision in complex converter station environments. By integrating invariant representation learning and a multi-task mixture of experts (MOE) framework, InvMOE demonstrated significant robustness and accuracy across diverse fault detection tasks. Our approach leverages invariant representation learning to disentangle task-relevant causal features from environmental noise, improving the model's generalization to out-of-distribution (OOD) scenarios. Additionally, the multi-task MOE framework enables adaptive routing of inputs to task-specific experts while effectively sharing knowledge across tasks. This design mitigates the impact of limited training samples and achieves improved performance for low-resource fault detection categories, such as valve cooling water leakage. Experimental results on real-world datasets confirm the efficacy of InvMOE. In future work, we aim to extend this framework by incorporating additional causal priors and exploring self-supervised pretraining strategies to further improve the robustness and scalability of fault detection systems in dynamic industrial environments.

Funding: This work was supported by the National Natural Science Foundation of China (No. 52167011).

 doi:10.20944/preprints202412.1342.v1

References

1. Jovcic, D.; Ahmed, K. *High Voltage Direct Current (HVDC) Transmission Systems*; Wiley, 2015.
2. Adamson, C.; Hingorani, N.G. *High-voltage direct-current power transmission*; Garraway, 1960.
3. Bu, S.; et al. Feature-based fault detection in converter stations. *IEEE Transactions on Industrial Electronics* **2017**, *64*, 7800–7808.
4. Sun, C.; et al. Hybrid methods for converter station monitoring. *Energy Reports* **2018**, *4*, 202–209.
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the NeurIPS, 2012, pp. 1097–1105.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR, 2016, pp. 770–778.
7. Gulrajani, I.; Lopez-Paz, D. In search of lost domain generalization. In Proceedings of the ICLR, 2021.
8. Arjovsky, M.; Bottou, L.; Gulrajani, I. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893* **2019**.
9. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the NeurIPS, 2015, pp. 91–99.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. Image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the ICLR, 2020.
11. Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research* **2018**, *19*, 1–34.
12. Li, Z.; et al. Deep learning-based fault diagnosis in HVDC systems. *IEEE Transactions on Power Electronics* **2019**.
13. Liu, W.; et al. Image-based fault detection in converter stations using deep learning. *IEEE Access* **2020**.
14. Li, Y.; et al. Advanced deep learning models for converter station fault detection. *IEEE Transactions on Industrial Informatics* **2021**.
15. Peng, C.; et al.. Out-of-Distribution Generalization: A Survey. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–35. https://doi.org/10.1145/3418076.
16. Pearl, J.; et al.. Causal Representation Learning: An Overview. *arXiv preprint arXiv:2001.09991* **2020**.
17. Creager, E.; Jacobsen, J.H.; Zemel, R. Environment inference for invariant learning. In Proceedings of the ICML. PMLR, 2021, pp. 2189–2200.
18. Krueger, D.; Caballero, E.; Jacobsen, J.H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In Proceedings of the ICML. PMLR, 2021, pp. 5815–5826.
19. Li, H.; Zhang, Z.; Wang, X.; Zhu, W. Learning invariant graph representations for out-of-distribution generalization. *NeurIPS* **2022**, *35*, 11828–11841.
20. Ajra, Y.; Hoblos, G.; Al Sheikh, H.; Moubayed, N. A Literature Review of Fault Detection and Diagnostic Methods in Three-Phase Voltage-Source Inverters. *Machines* **2024**, *12*, 631.
21. Wu, F.; Chen, K.; Qiu, G.; Zhou, W. Robust Open Circuit Fault Diagnosis Method for Converter Using Automatic Feature Extraction and Random Forests Considering Nonstationary Influence. *IEEE Transactions on Industrial Electronics* **2024**.
22. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering* **2021**, *34*, 5586–5609.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.