An Explainable Deep Learning Approach for Electrical Submersible Pump (ESP) Troubleshooting Using Ammeter Chart Image Interpretation

Yasmin Farisah Salma
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
yasminfsalma@gmail.com

Abstract*—Electrical Submersible Pump (ESP) is widely used as an artificial lift method in oil production, and its operating condition is commonly monitored through ammeter charts. Manual interpretation of ammeter charts depends heavily on engineer experience, making the diagnosis process subjective, time consuming, and difficult to standardize. This paper proposes an explainable deep learning approach for Electrical Submersible Pump troubleshooting through ammeter chart image interpretation. The main challenge addressed in this work is the acquisition gap between clean digital chart images and real camera images captured from field screens using mobile phones. To improve robustness, an acquisition augmentation strategy is designed to simulate geometric, photometric, and sensor-related degradations, including perspective distortion, brightness variation, glare, blur, noise, and image compression. Four convolutional neural network architectures are evaluated, namely SmallCNN, MobileNetV3, ResNet-50, and ConvNeXt-Tiny. Grad-CAM and LIME are integrated to provide visual explanations of model predictions. Experimental results show that ConvNeXt-Tiny achieves perfect performance on clean chart images and the best performance on real camera images, with an accuracy of 93.3 percent. Acquisition augmentation significantly improves robustness on camera images and reduces calibration error, indicating more reliable confidence estimation. The resulting prototype is implemented as a mobile-based diagnostic system that integrates image input, classification, and explainable visualization. These results indicate that the proposed approach can support more robust, transparent, and practical Electrical Submersible Pump troubleshooting in field conditions.

Keywords—acquisition gap; ammeter chart; convolutional neural network; deep learning; Electrical Submersible Pump; explainable artificial intelligence.

I. INTRODUCTION

Electrical Submersible Pump (ESP) is one of the most widely used artificial lift methods in oil production, especially for wells whose reservoir pressure is no longer sufficient to lift fluids naturally to the surface [1]. ESP systems are capable of supporting high production rates and are commonly deployed in mature oil fields. However, their operation is highly affected by downhole and surface conditions. Operational disturbances such as gas locking, pump-off, overload, undercurrent load, debris, and excessive cycling may reduce pump performance, increase downtime, and potentially damage expensive equipment. Therefore, early and reliable diagnosis of ESP operating conditions is essential to maintain production efficiency and support preventive maintenance decisions.

In field operations, ESP conditions are commonly monitored using ammeter charts, which record the electric current consumed by the ESP motor over time [1]. The current pattern in an ammeter chart reflects the motor load and provides diagnostic information regarding the operating state of the pump. Under normal conditions, the current tends to remain stable around its expected operating range. In contrast, faulty or abnormal conditions may produce characteristic patterns such as current spikes, gradual decreases, irregular fluctuations, repeated start-up cycles, or prolonged zero-current periods. These visual patterns are typically interpreted by engineers to identify the corresponding ESP condition.

Although manual ammeter chart interpretation has long been used in practice, it has several limitations. First, the interpretation process depends heavily on the experience and domain knowledge of the engineer. Similar current patterns may lead to different conclusions when interpreted by different individuals, especially for borderline cases between two operating conditions. Second, manual interpretation is time consuming and difficult to standardize when a large number of wells must be monitored. Third, the subjectivity of manual analysis may reduce consistency in operational decision-making. These limitations motivate the development of an automatic diagnostic system that can interpret ammeter chart patterns more consistently and efficiently.

Recent advances in machine learning and deep learning have enabled automatic fault diagnosis in industrial systems. Several studies have applied machine learning methods to ESP monitoring using numerical sensor data, including pressure, current, voltage, flow rate, and Supervisory Control and Data Acquisition data. Guo et al. applied machine learning to ESP monitoring using SCADA data for failure prediction [2]. Li et

al. used wellhead electrical parameters and production parameters for ESP condition monitoring and fault diagnosis [3]. However, most existing approaches focus on tabular or time-series data, while the use of ammeter chart images as the primary diagnostic input remains relatively underexplored. In practice, ammeter charts contain rich visual information that can be naturally formulated as an image classification problem. Convolutional Neural Networks (CNNs) are suitable for image classification tasks because they can learn hierarchical visual features without handcrafted feature engineering [4].

Despite this potential, image-based ESP diagnosis faces an important challenge when applied in real field conditions. Most classification models are trained using clean digital ammeter chart images, which usually have a white background, clear contrast, and minimal visual disturbance. However, in practical field scenarios, engineers may acquire chart images by taking photos of screen displays using mobile phone cameras. This difference in acquisition process creates an acquisition gap between the training data and the data encountered during deployment. Real camera images may suffer from perspective distortion, lighting variation, screen glare, blur, sensor noise, and compression artifacts. As shown in Fig. 1, the visual characteristics of a clean digital chart can differ significantly from both an acquisition-augmented image and a real camera image. Consequently, a model trained only on clean chart images may experience substantial performance degradation when tested on real camera images, even producing incorrect predictions with high confidence.

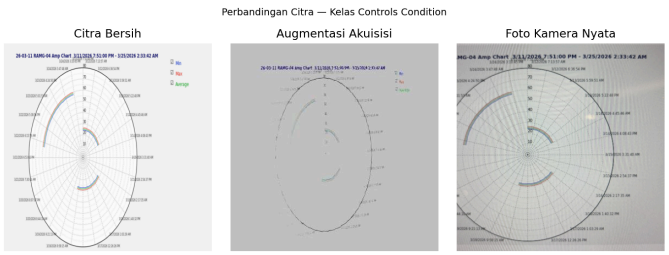

Fig. 1. Comparison between a clean digital ammeter chart, an acquisition-augmented image, and a real camera image captured from a screen display.

Another challenge is the limited interpretability of deep learning models. Deep learning models are often difficult to interpret because their decision-making process is commonly treated as a black-box mechanism [5]. In industrial diagnosis, especially in systems involving costly equipment and operational risks, engineers need more than a predicted class label. They also need to understand which parts of the input image influence the model prediction. Without interpretability, a model may rely on irrelevant visual cues such as chart text, grid lines, borders, or background artifacts instead of the current curve pattern that is meaningful from a domain perspective. Therefore, explainability is an important requirement for increasing trust and supporting human verification of model outputs.

To address these challenges, this paper proposes an explainable deep learning approach for ESP troubleshooting through ammeter chart image interpretation. The proposed method classifies ammeter chart images into fifteen ESP operating conditions and improves robustness against real camera acquisition through a targeted acquisition augmentation strategy. The augmentation process is designed to simulate geometric, photometric, and sensor-related degradations that commonly occur when charts are captured using mobile phone cameras. Four CNN architectures are evaluated, namely SmallCNN, MobileNetV3, ResNet-50, and ConvNeXt-Tiny, to analyze the effect of model capacity and architectural design on classification performance, robustness, and inference efficiency. In addition, Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated to provide class-discriminative visual explanations of CNN predictions [6]. Local Interpretable Model-agnostic Explanations (LIME) is also used to explain individual predictions through local input perturbation [7].

The main contributions of this paper are summarized as follows. First, this paper develops an image-based ESP troubleshooting approach that classifies ammeter chart images into fifteen operating conditions. Second, it proposes an acquisition augmentation strategy to reduce the distribution gap between clean digital chart images and real mobile camera images. Third, it presents a comparative evaluation of four CNN architectures on both clean test images and real camera test images. Fourth, it integrates Grad-CAM and LIME to improve prediction transparency by visualizing image regions that contribute to the classification result. Finally, it implements a mobile-based prototype to demonstrate an end-to-end diagnostic workflow, including image input, classification, and explainable visualization.

The remainder of this paper is organized as follows. Section II reviews related work on ESP fault diagnosis, deep learning for signal and chart image interpretation, and explainable artificial intelligence in industrial applications. Section III describes the proposed method, including dataset preparation, preprocessing, acquisition augmentation, CNN architectures, and explainability methods. Section IV presents the experimental setup and evaluation metrics. Section V discusses the experimental results, including classification performance, robustness on camera images, calibration, latency, and explainability analysis. Finally, Section VI concludes the paper and outlines future research directions.

II. Related Work

A. Electrical Submersible Pump (ESP) Fault Diagnosis

Fault diagnosis in Electrical Submersible Pump (ESP) systems is an important task in oil production because ESP failures may reduce production efficiency, increase maintenance cost, and cause operational downtime [1]. ESP operating conditions are commonly monitored using motor current, pressure, temperature, production parameters, and ammeter charts. In conventional field practice, engineers identify abnormal operating conditions by comparing observed operational patterns with known fault signatures, such as pump-off, gas lock, overload, undercurrent load, and

excessive cycling. Although this approach is widely used, it depends heavily on engineer experience and may lead to inconsistent interpretation when different fault conditions produce visually similar patterns.

Data-driven approaches have been developed to reduce the dependency on manual interpretation. Guo et al. applied Principal Component Analysis (PCA) and XGBoost to SCADA data for predicting ESP failures several days in advance [2]. This study shows that machine learning can capture nonlinear relationships among operational variables and support early warning systems for ESP failure prediction. Li et al. utilized wellhead electrical parameters and production parameters for ESP condition monitoring and fault diagnosis [3]. These studies demonstrate the potential of machine learning for identifying abnormal ESP behavior from operational data.

Recent studies have further explored deep learning and hybrid approaches for ESP fault diagnosis. Gong et al. proposed a fault diagnosis model that combines mechanism knowledge and deep learning for ESP systems [8]. Song et al. introduced a deep learning model for ESP failure diagnosis using experimental sand-water flow data [9]. Yang et al. combined unsupervised learning and multi-source transfer learning to improve ESP fault identification [10]. These studies indicate that deep learning can improve ESP diagnosis by learning complex patterns from sensor data, experimental data, or transferred domain knowledge. Despite these advances, most ESP fault diagnosis studies are still dominated by numerical sensor data, SCADA records, or time-series signals. These data sources are useful for capturing operational behavior, but they do not directly exploit the visual information contained in ammeter charts. In field operations, ammeter charts remain important because their current patterns provide intuitive visual indications of ESP operating conditions. However, the use of ammeter chart images as the primary input for automated ESP troubleshooting remains relatively limited.

Another limitation of existing studies is the lack of robustness evaluation under camera-based acquisition conditions. In practical scenarios, engineers may capture ammeter charts by photographing screen displays using mobile phone cameras. This acquisition process can introduce perspective distortion, glare, illumination changes, blur, noise, and compression artifacts. Therefore, a model trained only on clean digital chart images may not generalize well to real camera-captured images. Furthermore, explainability remains an important issue for ESP fault diagnosis. In industrial environments, engineers need to verify whether a model bases its decision on meaningful operational patterns rather than irrelevant artifacts. Existing ESP diagnosis studies have not widely investigated explainable AI for ammeter chart image interpretation. Therefore, this paper focuses on image-based ESP troubleshooting using ammeter chart classification, acquisition augmentation for camera robustness, and visual explanation using Grad-CAM and LIME.

B. Deep Learning for Signal and Chart Image Interpretation

Deep learning has been widely applied to signal interpretation by transforming one-dimensional time-series data into two-dimensional image representations. This approach enables temporal patterns in signals to be represented as spatial structures that can be processed by computer vision models. Common time-series-to-image representations include spectrograms, scalograms, Gramian Angular Field (GAF), Markov Transition Field (MTF), and recurrence plots. These representations are useful because they can capture different characteristics of the original signal, such as time-frequency distribution, temporal correlation, state transition patterns, and repeated structures.

Convolutional Neural Networks (CNNs) are commonly used for image-based signal interpretation because they can automatically learn hierarchical visual features from input images [4]. Through convolutional layers, CNNs can extract local patterns such as edges, curves, textures, and repeated structures. These features are then combined into higher-level representations that are useful for classification. This capability makes CNNs suitable for analyzing visual representations of signals without requiring manual feature engineering. Wang and Oates introduced Gramian Angular Field and Markov Transition Field as image representations for time-series classification [11]. Their work shows that temporal information can be encoded into image-like structures that are suitable for deep learning models. This idea has influenced many later studies that use visual representations to analyze signals in different domains. Instead of relying only on raw numerical sequences, the signal is represented as an image so that spatial feature extraction can be applied.

In industrial applications, image-based signal interpretation has been used for machine and motor fault diagnosis. Valtierra-Rodríguez et al. applied CNN and motor current signature analysis for detecting broken rotor bars in induction motors [12]. In this type of approach, fault-related patterns in the signal can appear as discriminative visual structures that are learned by the CNN. These results show that current-related signals can contain diagnostic information that is suitable for CNN-based analysis. Image-based signal interpretation has also been applied in medical signal analysis. Zhang et al. used recurrence plot-based image representations for cardiac arrhythmia classification [13]. This demonstrates that signal-to-image representation can preserve important diagnostic information from temporal signals. The success of this approach in different domains supports the idea that visual patterns derived from signals can be effectively analyzed using deep learning.

This paper is related to these studies because ammeter charts also represent current behavior over time. However, the problem addressed in this work is different from conventional time-series-to-image transformation. In this study, the input is not a raw numerical signal that is converted into an image, but an ammeter chart that already exists as a visual chart used in field operations. Therefore, the main task is to interpret existing chart images rather than generate image representations from raw signals. This distinction introduces additional challenges. Previous signal-to-image studies usually assume that the generated image representation is clean and controlled. In practical ESP monitoring, ammeter chart images may be acquired by photographing screen displays using

mobile phone cameras. As a result, the images may contain perspective distortion, glare, blur, lighting variation, noise, and compression artifacts. These acquisition-related degradations can reduce model performance when the model is trained only on clean digital chart images. Furthermore, many previous studies focus mainly on classification performance and do not explicitly analyze whether the model attends to physically meaningful signal regions. In ESP troubleshooting, it is important to verify whether the model focuses on the current curve pattern rather than irrelevant chart elements. Therefore, this paper extends image-based signal interpretation to the ESP ammeter chart domain by combining CNN-based classification, acquisition augmentation, and explainable visualization.

C. *Explainable Artificial Intelligence (XAI) in Industrial Diagnosis*

The deployment of deep learning models in industrial diagnosis requires not only high predictive performance but also interpretability. Industrial systems often involve costly equipment, safety considerations, and operational decisions that must be justified by engineers. Although deep learning models can learn complex patterns from visual or sensor data, their internal decision-making process is often difficult to understand. Deep learning models are often treated as black-box systems because users cannot directly observe how the model transforms input features into final predictions [5].

Explainable Artificial Intelligence (XAI) has therefore become an important research direction in predictive maintenance and industrial condition monitoring. XAI methods aim to provide explanations that help users understand, verify, and audit model predictions. Hosain et al. reviewed recent advancements, applications, and challenges of explainable AI in deep learning [14]. Their study emphasizes that explainability is important for increasing trust and supporting the practical adoption of deep learning models. Moosavi et al. also surveyed explainable AI applications in manufacturing and industrial cyber-physical systems [15]. Their survey highlights that transparent AI models are needed when predictions are used to support operational decisions in industrial environments.

For image-based diagnosis, Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely used post-hoc explanation methods [6]. Grad-CAM produces a class-discriminative heatmap by using gradient information from the last convolutional layer of a CNN. The heatmap highlights image regions that contribute strongly to a target class prediction. In industrial applications, this type of visualization can help engineers verify whether the model focuses on relevant visual evidence, such as surface defects, thermal hotspots, vibration-related patterns, or signal structures in time-frequency images. Explainability has also been applied in industrial condition monitoring to relate model predictions to physical or operational patterns. Brusa et al. used explainable AI to analyze feature contributions in machine learning models for industrial condition monitoring [16]. Their work shows that explanation methods can help users understand which input regions or features influence the prediction. This is important because a model with high accuracy may still be unreliable if it bases its decision on irrelevant artifacts.

Another commonly used XAI method is Local Interpretable Model-agnostic Explanations (LIME) [7]. LIME explains individual predictions by perturbing parts of the input and approximating the model behavior locally using an interpretable surrogate model. For image data, LIME usually divides the image into superpixels and identifies regions that support or oppose a specific prediction. This makes LIME useful for inspecting whether a prediction depends on relevant chart components rather than background elements or visual noise. Although XAI methods have been increasingly used in industrial diagnosis, their application to ammeter chart-based ESP troubleshooting remains limited. Most existing XAI studies focus on physical images, thermal images, tabular sensor data, or transformed signal images. Direct explainability for operational chart images, such as ESP ammeter charts, has not been widely investigated. In this paper, Grad-CAM is used to provide heatmap-based explanations of CNN predictions. LIME is also used to provide superpixel-based local explanations of individual predictions. The objective is to verify whether the model focuses on current curve patterns that are meaningful for ESP diagnosis rather than irrelevant elements such as labels, borders, grid lines, or screen artifacts.

D. *Research Gap*

Based on the reviewed studies, several research gaps can be identified as follows:

1) Limited use of ammeter chart images for ESP fault diagnosis
   Most existing ESP fault diagnosis approaches rely on numerical sensor data, SCADA records, or time-series signals. Although these data sources are useful for capturing operational behavior, they do not directly exploit the visual diagnostic information contained in ammeter charts. In field practice, ammeter charts are widely used by engineers because their current patterns provide intuitive visual indications of ESP conditions. However, the formulation of ESP troubleshooting as an image classification problem using ammeter chart images remains relatively limited.

2) Lack of robustness evaluation under camera-based acquisition conditions
   Previous studies on signal and chart interpretation generally assume that the visual representation used for training and testing is clean, controlled, or synthetically generated from numerical signals. This assumption differs from real field scenarios, where engineers may acquire ammeter chart images by photographing screen displays using mobile phone cameras. Such images may contain perspective distortion, glare, illumination changes, blur, noise, and compression artifacts. Therefore, the acquisition gap between clean digital chart images and real camera-captured images becomes an important practical issue. Existing ESP diagnosis studies rarely evaluate model robustness under this type of acquisition shift.

3) Limited application of XAI for ammeter chart-based ESP diagnosis
   Although deep learning models have shown strong performance in industrial diagnosis, their limited interpretability remains a barrier to field adoption. For ESP troubleshooting, engineers need to verify whether a model bases its prediction on meaningful current curve patterns rather than irrelevant visual elements such as chart labels, borders, grid lines, or screen artifacts. While XAI methods such as Grad-CAM and LIME have been applied in other industrial domains, their use for explaining CNN-based ammeter chart interpretation in ESP diagnosis is still underexplored.
4) Insufficient joint evaluation of robustness, calibration, explainability, and deployment feasibility
   Many existing studies focus mainly on classification accuracy. However, in practical diagnostic systems, high accuracy alone is insufficient. The model must also remain reliable under real acquisition conditions, provide confidence estimates that are consistent with actual correctness, and produce explanations that can be inspected by users. In addition, field-oriented deployment requires an end-to-end workflow that supports image input, classification, and result interpretation through an accessible interface.

III. METHODOLOGY

This section describes the proposed methodology for explainable Electrical Submersible Pump (ESP) troubleshooting using ammeter chart image interpretation. The methodology is designed to address three main requirements: accurate classification of ESP operating conditions, robustness against real camera acquisition, and interpretability of model predictions. The proposed workflow consists of dataset preparation, image preprocessing, acquisition augmentation, CNN-based classification, explainability generation, and mobile-based system integration. The overall system architecture is illustrated in Fig. 2.

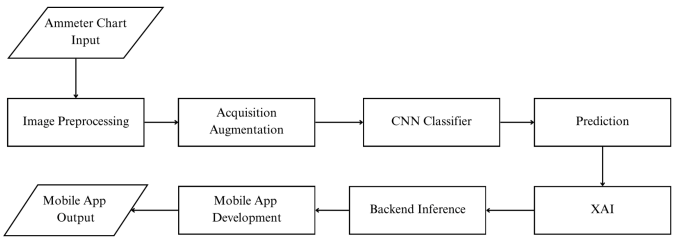

Fig. 2. Overview of the proposed explainable deep learning system for ammeter chart-based ESP troubleshooting.

A. System Overview

The proposed system is developed as an end-to-end diagnostic workflow for classifying ammeter chart images into Electrical Submersible Pump (ESP) operating conditions and providing visual explanations for the resulting predictions. As shown in Fig. 2, the system consists of two main components: a machine learning pipeline and a mobile-based user interface. These two components are connected through an inference backend that receives input images, performs model inference, generates explainability outputs, and returns the diagnostic results to the user. The machine learning pipeline is responsible for training and evaluating the image classification model. The input data consist of clean digital ammeter chart images and real camera-captured ammeter chart images. Clean digital images are used as the main training data, while real camera images are used to evaluate model robustness under acquisition conditions that commonly occur in field scenarios. Before being processed by the model, each image undergoes preprocessing to isolate the chart region, standardize the input size, and normalize the pixel values. This preprocessing step ensures that the model receives consistent input and focuses on the current curve pattern rather than irrelevant visual elements.

During training, acquisition augmentation is applied to simulate visual degradation observed in real camera images. The augmentation process includes geometric, photometric, and sensor-related transformations, such as perspective distortion, brightness variation, glare, blur, noise, and compression artifacts. This strategy is intended to reduce the distribution gap between clean digital charts and camera-captured charts. A clean-mix strategy is also applied by retaining a portion of the training images in their clean form. This allows the model to learn both ideal chart distributions and degraded camera-like chart distributions.

The classification module is built using Convolutional Neural Network (CNN) architectures. Four models are evaluated in this study, namely SmallCNN, MobileNetV3, ResNet-50, and ConvNeXt-Tiny. MobileNetV3 is selected as an efficient CNN architecture for resource-constrained scenarios [17]. ResNet-50 is selected as a classical residual CNN baseline with strong representational capacity [18]. ConvNeXt-Tiny is selected as a modern CNN architecture with updated convolutional design [19]. SmallCNN is used as a lightweight custom baseline developed in this study.

The training strategy differs between the custom model and the pretrained models. SmallCNN is trained from scratch using only the ammeter chart dataset. The pretrained models use ImageNet-pretrained weights to initialize general visual representations [20]. Transfer learning is used to adapt pretrained visual representations to the ammeter chart classification task [21]. The comparison among these models is used to analyze the trade-off between classification accuracy, robustness to camera images, and inference efficiency. After a trained model produces a prediction, the explainability module generates visual explanations. Grad-CAM is used to highlight image regions that contribute strongly to the predicted class based on gradient information from the CNN [6]. LIME is used to identify influential superpixel regions through local input perturbation [7]. These explanations allow users to verify whether the model focuses on meaningful ammeter chart patterns, especially the current curve, rather than irrelevant components such as labels, borders, grid lines, or background artifacts.

The mobile application serves as the user-facing component of the system. It allows users to capture or select an ammeter chart image, crop the relevant chart area, submit the image to the backend, and receive the diagnostic output. The final output includes the predicted ESP condition, confidence score, class probability distribution, and visual explanation results. By integrating classification and explainability into a mobile-based workflow, the proposed system is intended to support practical field usage and provide engineers with a more transparent diagnostic aid.

B. Dataset

The dataset used in this study consists of ammeter chart images representing fifteen Electrical Submersible Pump (ESP) operating conditions. The classes include normal condition, power fluctuations, gas locking, pump-off, false starts, excessive cycling, gassy condition, undercurrent load, UC below no load, controls condition, overload, debris, excess restarts, erratic pattern, and stuck close. These classes were selected based on common ESP operating patterns observed in ammeter charts. Each image was labeled according to the corresponding current pattern and operating condition. To reduce individual subjectivity in the labeling process, the annotation was conducted by six annotators using standardized class descriptions as guidance.

The dataset contains two types of images: clean digital ammeter chart images and real camera-captured ammeter chart images. The clean dataset consists of 7,098 images collected from field operational data and secondary sources from PT Pertamina Hulu Energi OSES. These images represent ideal chart conditions, with clear contrast, minimal visual noise, and relatively consistent digital rendering. The clean images were divided into training, validation, and test sets using a stratified 80:10:10 split, resulting in 5,676 training images, 711 validation images, and 711 test images. The stratified split was applied to preserve the class distribution across all subsets. A fixed random seed of 42 was used to ensure reproducibility.

In addition to the clean dataset, a real camera dataset was collected to evaluate model robustness under practical acquisition conditions. This dataset consists of 150 images captured by photographing ammeter charts from screen displays using a mobile phone camera. Each class contains 10 camera images. The camera dataset was divided into two disjoint subsets: a camera validation set and a camera test set. The camera validation set contains 75 images, with 5 images per class, and was used for model selection and augmentation calibration. The camera test set also contains 75 images, with 5 images per class, and was kept as a held-out test set for final robustness evaluation. This separation ensures that the final evaluation on camera images remains independent from the model selection process.

The camera images represent acquisition conditions that are commonly encountered in field scenarios. Compared with clean digital images, these images may contain perspective distortion, brightness variation, glare from screen reflection, blur, sensor noise, and compression artifacts. Therefore, the camera dataset is used to measure the ability of the model to generalize from clean chart images to real mobile camera images. The overall dataset composition is summarized in Table I.

TABLE I. DATASET COMPOSITION

Dataset Subset	***Image Type***	***Number of Images***	***Number of Classes***	***Images per Class***	***Usage***
Training set	Clean digital chart	5,676	15	358-399	Model training
Validation set	Clean digital chart	711	15	45-50	Clean validation
Clean test set	Clean digital chart	711	15	45-50	Evaluation on ideal images
Camera validation set	Real camera image	75	15	5	Model selection and augmen-tation calibration
Camera test set	Real camera image	75	15	5	Final robustness evaluation
Total	Clean and camera images	7,248	15	-	Full experimental

Before being used for model training and evaluation, all images were processed using the same preprocessing pipeline. The preprocessing includes chart region cropping, resizing to 224 × 224 pixels, conversion to RGB format, and normalization using ImageNet mean and standard deviation. This ensures that all image subsets share a consistent input format while preserving the visual current pattern that is relevant for ESP condition classification.

C. Image Preprocessing

Image preprocessing is applied to transform raw ammeter chart images into consistent inputs for the classification model. This stage is designed to reduce irrelevant visual information while preserving the chart region that contains the current curve pattern. The preprocessing pipeline consists of the following steps:

1) Chart Region Cropping
   The first step is to isolate the main ammeter chart region from non-signal elements such as text, borders, labels, axes, and background areas. This process is implemented using contour detection based on OpenCV. Each image is converted into grayscale and processed using thresholding to separate the chart structure from the background. The largest contour that resembles the circular chart area is then selected as the candidate region, and the image is cropped based on its bounding box.

2) Image Resizing
   After the chart region is cropped, the resulting image is resized to 224 × 224 pixels. This size is used to match the input requirement of the CNN architectures evaluated in this study, especially the ImageNet-pretrained models. Resizing also ensures that all images have consistent

dimensions during training, validation, testing, and inference.

3) Image Normalization
The pixel values are normalized using the ImageNet mean and standard deviation to align the input distribution with the ImageNet pretraining setting [20]. The same normalization procedure is applied consistently to clean digital images, acquisition-augmented images, and real camera images.

Through these preprocessing steps, the input images are standardized while retaining the current curve pattern as the main diagnostic information for ESP condition classification.

D. Acquisition Augmentation

Acquisition augmentation is designed to reduce the acquisition gap between clean digital ammeter chart images and real camera-captured images. In the training data, clean chart images have ideal visual quality, while real field images captured using mobile phone cameras may contain several types of degradation caused by screen photographing. These degradations include perspective distortion, lighting variation, screen glare, blur, sensor noise, and compression artifacts. Therefore, the augmentation strategy is specifically designed to simulate the degradation patterns observed in real camera images.

The augmentation is applied on-the-fly during training to increase data variation and improve model generalization [22]. This allows different augmented versions of the same image to be generated across training epochs without physically increasing the dataset size. Through this process, the model is exposed to diverse acquisition conditions and is encouraged to learn current curve patterns that remain stable under visual degradation. This strategy is related to domain randomization, where synthetic variations are introduced during training to improve robustness under target-domain shifts [23]. The augmentation methods are grouped into three categories according to the source of degradation they simulate, as summarized in Table II.

TABLE II. Acquisition Augmentation Groups and Simulated Degradation

Group	***Augmentation Methods***	***Simulated Degradation***
Geometric	Perspective warp, small rotation, zoom	Camera angle and position variation
Photometric	Brightness/contrast adjustment, gamma, glare, screen-dim	Lighting variation, screen reflection, and display characteristics
Sensor degradation	Gaussian blur, JPEG compression, noise	Camera focus, image compression, and sensor characteristics

Two design decisions are applied in the acquisition augmentation strategy:

1) Calibration Based on Real Camera Images
The augmentation parameters are adjusted based on the observed characteristics of the collected camera images. For example, the cropped camera images have an average brightness level of approximately 0.77 compared with clean digital images. Therefore, the screen-dim transformation is designed to generate a comparable brightness reduction so that the augmented images remain realistic.

2) Clean-Mix Strategy
A portion of the training images is kept in clean form without augmentation. In this study, approximately 20% of the training data is preserved as clean images. This strategy prevents the model from becoming overly biased toward degraded images and helps maintain performance on clean digital charts. Acquisition augmentation is applied only to the training set, while the validation and test sets remain unchanged.

Through this strategy, the model is expected to learn both the ideal distribution of clean ammeter chart images and the degraded distribution that resembles real camera acquisition conditions.

E. Convolutional Neural Network (CNN) Architecture and Training Strategy

This study evaluates four Convolutional Neural Network (CNN) architectures to analyze the effect of model capacity, architectural design, and transfer learning on ammeter chart classification. The evaluated models consist of SmallCNN, MobileNetV3-Large, ResNet-50, and ConvNeXt-Tiny. These models were selected to represent different levels of complexity, ranging from a lightweight custom CNN to larger ImageNet-pretrained architectures. The comparison of the evaluated architectures is summarized in Table III.

TABLE III. Comparison of Evaluated CNN Architectures

Model	***Architecture***	***Parameters***	***Era***	***Training Paradigm***
SmallCNN	SmallCNN	2.8M	Custom	Architecture developed in this study
MobileNetV3	MobileNetV3 -Large	4.2M	Efficient CNN, 2019	Transfer Learning
ResNet-50	ResNet-50	23.5M	Classical residual CNN, 2016	Transfer Learning
ConvNeXt -Tiny	ConvNeXt -Tiny	28M	Modern CNN, 2022	Transfer Learning

The role of each model is described as follows:

1) SmallCNN
SmallCNN is a lightweight CNN architecture developed in this study for ammeter chart image classification. It consists of a stem, four feature extraction stages, and a classification head. The model is designed to capture current curve patterns such as stable curves, current spikes, decreasing current patterns, start-stop cycles, and repeated fluctuations while maintaining a relatively small number of parameters.

2) MobileNetV3-Large
MobileNetV3-Large is selected as an efficient CNN architecture for resource-constrained scenarios [17]. It uses a mobile-oriented architectural design that makes it relevant for practical diagnostic systems requiring relatively low computational cost. In this study, MobileNetV3-Large represents an efficient transfer learning model for evaluating the balance between speed and robustness.

3) ResNet-50
ResNet-50 is selected as a classical residual CNN baseline with strong representational capacity [18]. Its residual learning mechanism allows deeper convolutional networks to be trained more effectively. In this study, ResNet-50 is used to evaluate whether a higher-capacity residual architecture can improve robustness against acquisition-related distribution shift.
4) ConvNeXt-Tiny
ConvNeXt-Tiny is selected as a modern CNN architecture with updated convolutional design [19]. It represents a newer generation of CNNs that incorporates several modern architectural choices while retaining the convolutional structure. In this study, ConvNeXt-Tiny is used to evaluate whether modern CNN design can improve performance and robustness for ammeter chart image classification.

The training strategy differs between the custom CNN and the pretrained models. SmallCNN is trained from scratch without pretrained weights, so all of its parameters are optimized directly using the ammeter chart dataset. The pretrained models use ImageNet-pretrained weights to initialize general visual representations [20]. Transfer learning is then used to adapt these pretrained visual representations to the ammeter chart classification task [21].

For the transfer learning models, the input image is first processed by the pretrained backbone to extract visual feature representations. These features are then summarized using global average pooling and passed to a classification head consisting of dropout, a linear layer, and softmax activation to produce probabilities for the fifteen ESP operating classes. The transfer learning process is performed using staged fine-tuning in two stages:

1) Frozen Backbone
Stage In the first stage, the pretrained backbone is frozen, and training is focused on the classification head. This stage adapts the output layer to the ESP condition labels without changing the general visual representation learned from ImageNet.
2) Full Fine-Tuning
Stage In the second stage, the entire network is unfrozen and fine-tuned. A smaller learning rate is used for the backbone than for the classification head. This allows the model to adapt more specifically to ammeter chart images while maintaining training stability and reducing the risk of overly drastic parameter changes.

Several training components are also applied to improve stability and robustness:

1) Combination of Clean and Augmented Images
The training process uses both clean images and acquisition-augmented images. This allows the model to learn from ideal chart images as well as images that resemble camera-captured field data.
2) Clean-Mix Training
A portion of the training images remains clean during training. This is intended to preserve model performance on clean digital images while still improving robustness against degraded camera-like images.
3) Dual Validation Strategy
At each epoch, the model is evaluated using both clean validation data and degraded validation data. The clean validation set measures performance under ideal conditions, while the degraded validation set measures robustness against acquisition-like degradation.
4) Checkpoint Selection
The best checkpoint is selected based on the average macro-F1 score from the clean validation set and the degraded validation set. This criterion ensures that the selected model performs well under both clean and degraded conditions.
5) Optimization Stability
Gradient clipping, smaller backbone learning rate, and full-precision training are applied to reduce training instability, especially when strong acquisition augmentation is used.
6) Early Stopping
Training is stopped when validation performance does not improve for a certain number of epochs. This reduces overfitting and ensures that the selected model is based on validation performance rather than the final epoch alone.

The training hyperparameters used in this study are summarized in Table IV.

TABLE IV. TRAINING HYPERPARAMETERS

Parameter	***Value***
Optimizer	AdamW
Learning rate for head/backbone	1×10^{-3} / 2×10^{-5} for transfer learning
Scheduler	Cosine annealing
Loss function	Cross-entropy + label smoothing
clean_prob	0.2
Gradient clipping	1
Model selection	Average macro-F1 of clean and degraded validation sets

The model selection criterion is not based only on clean validation accuracy. Instead, the average macro-F1 score from clean and degraded validation data is used because the objective of this study is not only to obtain high accuracy on ideal images, but also to select a model that remains reliable under acquisition degradation.

F. Explainability Methods

Explainability is integrated into the proposed system to provide visual evidence for the CNN prediction. Since ESP troubleshooting is a domain-specific diagnostic task, the model output should not only provide the predicted class, but also indicate which image regions contribute to the decision. In this study, two post-hoc explainability methods are used, namely Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME).

1) Grad-CAM
Grad-CAM is used to generate a class-discriminative heatmap that highlights the regions of an ammeter chart image that contribute most strongly to the predicted class.

This method uses the gradient of the target class score with respect to the feature maps of the last convolutional layer. For a target class c, the importance weight of feature map k is computed using the Grad-CAM formulation [6].

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

where y^c denotes the score for class c, A_{ij}^k denotes the activation value at spatial location (i, j) in feature map k, and Z denotes the total number of pixels in the feature map. The Grad-CAM localization map is then obtained by computing the weighted combination of the feature maps, followed by a ReLU operation:

$$L_{Grad-CAM}^c = RelU\left(\sum_k \alpha_k^c A^k\right) \quad (2)$$

The resulting localization map is upsampled to the original image size and overlaid on the input ammeter chart. In this study, the heatmap is used to verify whether the model focuses on the current curve pattern, such as stable regions, spikes, drops, fluctuations, or repeated cycles, instead of irrelevant chart components.

2) LIME

LIME is used as a model-agnostic explanation method to identify local image regions that support the model prediction. Unlike Grad-CAM, which relies on internal gradients of the CNN, LIME explains an individual prediction by optimizing a local interpretable surrogate model around the input instance [7]. Formally, LIME searches for an interpretable model g that approximates the original model f around an input instance x. The optimization objective is defined as

$$\xi(x) = arg\,min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (3)$$

where f denotes the original classification model, g denotes the interpretable surrogate model, G denotes the class of interpretable models, π_x defines the locality or proximity around instance x, $L(f, g, \pi_x)$ measures the fidelity of the surrogate model to the original model in the local region, and $\Omega(g)$ represents the complexity penalty of the explanation model. For image data, LIME first divides the input ammeter chart into superpixels. Several perturbed versions of the image are then generated by turning selected superpixels on or off. Each perturbed image is passed through the trained CNN model, and the prediction changes are used to estimate which superpixels contribute positively or negatively to the predicted class. The final explanation is visualized by highlighting the superpixel regions that most strongly support the prediction.

The two methods provide complementary forms of explanation. Grad-CAM produces a continuous heatmap based on the internal feature representation of the CNN, while LIME produces a local superpixel-based explanation by observing model behavior under input perturbations. In the proposed system, both explanations are generated after the model predicts an ESP condition. The explanations are then displayed together with the predicted label and confidence score to help users inspect whether the prediction is based on meaningful ammeter chart patterns. For ESP troubleshooting, the expected explanation should highlight the current curve region rather than unrelated elements such as chart labels, borders, grid lines, or background artifacts. Therefore, the XAI outputs are used not only as visualization tools, but also as a qualitative verification mechanism for assessing whether the model decision is aligned with domain-relevant visual patterns.

IV. Experimental Setup

This section describes the experimental setup used to evaluate the proposed explainable deep learning approach for ESP troubleshooting. The evaluation is designed to measure model performance under clean digital image conditions and real camera acquisition conditions. In addition, the experiments assess the effect of acquisition augmentation, calibration quality, inference efficiency, and explainability relevance.

A. Implementation Environment

The system is implemented using several development environments according to the requirements of model training, evaluation, inference deployment, and mobile application development. The machine learning pipeline is developed in Python using PyTorch as the main deep learning framework. The timm library is used to load ImageNet-pretrained CNN architectures, OpenCV is used for image preprocessing, and scikit-learn is used for metric computation. Model training is performed using NVIDIA A100 or T4 GPUs through Google Colab. Model evaluation, explainability analysis, and application testing are conducted on Apple Silicon using MPS and CPU execution.

The inference backend is implemented using FastAPI and deployed using Docker on HuggingFace Spaces. The mobile application is developed for iOS using SwiftUI, Xcode, and Clean Architecture. Grad-CAM and LIME explanations are generated using pytorch-grad-cam and LIME libraries. The implementation environment is summarized in Table V.

TABLE V. Implementation and Testing Environment

Component	***Environment***
Machine learning pipeline	Python 3.11, PyTorch, timm, OpenCV, scikit-learn
Model training	NVIDIA A100/T4 GPU through Google Colab
Evaluation and XAI	Apple Silicon, MPS, CPU
Inference backend	FastAPI, Docker, HuggingFace Spaces
Mobile application	SwiftUI, iOS, Xcode, Clean Architecture
Explainability methods	Pytorch-grad-cam, LIME

B. Evaluation Protocol

The evaluation protocol is designed to compare the four CNN architectures under the same experimental setting. The evaluated models are SmallCNN, MobileNetV3, ResNet-50, and ConvNeXt-Tiny. Each model is trained using the training strategy described in Section III, including preprocessing, acquisition augmentation, clean-mix training, and checkpoint selection based on validation performance. The experiments are conducted using two types of test data:

1) Clean test set
   The clean test set consists of digital ammeter chart images with balanced distribution across the fifteen ESP operating classes. This test set is used to evaluate model performance under ideal image conditions. Evaluation on clean images also serves as a regression test to ensure that acquisition augmentation does not reduce model performance on clean digital charts.
2) Camera test set
   The camera test set consists of real ammeter chart images captured from screen displays using a mobile phone camera. This set is used as the final held-out test set for evaluating robustness against real acquisition conditions. The camera test set is not used during model training or model selection.

To evaluate the effectiveness of acquisition augmentation, each architecture is compared under two training conditions:

1) Baseline training
   The model is trained using clean digital images only, without acquisition augmentation. This condition represents the standard training approach and serves as the baseline.
2) Augmented training
   The model is trained using acquisition augmentation that simulates geometric, photometric, and sensor-related degradations. This condition is used to evaluate whether the augmentation strategy improves robustness on real camera images.

The comparison between baseline and augmented models is used to measure the impact of acquisition augmentation on camera robustness. In addition, model calibration is evaluated on camera images to assess whether the confidence score produced by the model is consistent with its actual correctness. Finally, inference latency is measured to evaluate the trade-off between model accuracy and computational efficiency. Explainability is evaluated qualitatively using Grad-CAM and LIME visualizations. The generated visual explanations are inspected to determine whether the model focuses on domain-relevant current curve regions rather than irrelevant chart components such as text, borders, grid lines, or background artifacts.

C. Evaluation Metrics

Several evaluation metrics are used to measure classification performance, robustness, calibration quality, statistical significance, and inference efficiency. The metrics used in this study are accuracy, macro-F1, Expected Calibration Error (ECE), McNemar test, and inference latency.

1) Accuracy
   Accuracy measures the proportion of correctly classified samples among all evaluated samples. It is used to provide an overall view of model classification performance on both clean and camera test sets. Accuracy is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where TP, TN, FP, FN denote true positive, true negative, false positive, and false negative, respectively.

2) Macro-F1
   Macro-F1 is used to evaluate classification performance across all fifteen ESP classes by giving equal weight to each class [24]. This metric is important because the task is formulated as a multiclass classification problem. Macro-F1 ensures that the evaluation reflects performance across all ESP operating conditions rather than being dominated by a specific class. Macro-F1 is defined as

$$Macro - F1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{5}$$

where N denotes the number of classes and $F1_i$ denotes the F1-score of class i.

3) Expected Calibration Error
   Expected Calibration Error is used to measure the mismatch between prediction confidence and actual accuracy [25]. A lower ECE indicates that the model confidence is more consistent with its real accuracy. In this study, ECE is important because a diagnostic system should not only make correct predictions but also provide reliable confidence estimates. ECE is defined as

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{6}$$

where M denotes the number of confidence bins, B_m denotes the set of predictions in bin m, n denotes the total number of predictions, $acc(B_m)$ denotes the accuracy of bin m, and $conf(B_m)$ denotes the average confidence of bin m.

4) McNemar Test
   The McNemar test is used to compare two classifiers evaluated on the same test set [26]. In this study, it is used to determine whether the performance difference between the baseline model and the augmented model is statistically significant. This test is appropriate because both models are evaluated on the same camera test set. The McNemar statistic is defined as

$$x^2 = \frac{(|b-c|-1)^2}{b+c} \quad (7)$$

where b denotes the number of samples correctly classified by the baseline model but incorrectly classified by the augmented model, and c denotes the number of samples incorrectly classified by the baseline model but correctly classified by the augmented model. A result is considered statistically significant when the p-value is less than 0.05.

5) Inference Latency

Inference latency measures the time required by a model to produce a prediction. This metric is used to analyze the efficiency of each architecture and its suitability for practical deployment in a mobile-based diagnostic workflow. Lower latency indicates faster prediction time, while higher accuracy under camera acquisition indicates better robustness. Therefore, latency is analyzed together with camera test accuracy to understand the trade-off between computational efficiency and diagnostic performance.

V. Results and Discussion

A. Performance in Clean Images

The first evaluation was conducted on the clean test set to measure the classification performance of each model under ideal digital chart conditions. This evaluation also serves as a regression test to ensure that the use of acquisition augmentation during training does not reduce model performance on clean ammeter chart images. The results are summarized in Table VI.

TABLE VI. Classification Performance on Clean Test Images

Model	*Parameters (M)*	*Accuracy (%)*	*Macro-F1*	*ECE*
SmallCNN	2.8	99.3	0.993	0.039
MobileNetV3	4.2	98.45	0.984	0.040
ResNet-50	23.5	98.87	0.989	0.048
ConvNeXt-Tiny	28	100	1	0.045

As shown in Table VI, all evaluated CNN architectures achieved high performance on clean ammeter chart images, with accuracy values above 95%. This result indicates that the visual patterns contained in ammeter charts can be effectively learned by CNN-based models. The high macro-F1 scores also show that the models performed consistently across the fifteen ESP operating classes, rather than only performing well on a subset of classes.

ConvNeXt-Tiny achieved the best performance, with an accuracy of 100.00% and a macro-F1 score of 1.000. This suggests that the modern CNN design and larger representational capacity of ConvNeXt-Tiny are effective for capturing discriminative ammeter chart patterns under clean image conditions. SmallCNN also achieved strong performance, with an accuracy of 99.30% and a macro-F1 score of 0.993, despite having only 2.8 million parameters. This indicates that the classification task on clean digital charts can be handled effectively even by a lightweight CNN when the input distribution is ideal and visually consistent.

MobileNetV3 and ResNet-50 also produced high accuracy scores of 98.45% and 98.87%, respectively. Although these results are slightly lower than ConvNeXt-Tiny and SmallCNN, they remain far above the minimum success criterion of 85% accuracy defined for clean image evaluation. The ECE values are also relatively low for all models, ranging from 0.039 to 0.048, indicating that the confidence scores are reasonably aligned with prediction correctness under clean image conditions.

Overall, the clean image evaluation confirms that all models successfully learned the characteristic current patterns of ammeter charts. More importantly, the results show that the acquisition augmentation and clean-mix training strategy did not degrade performance on ideal digital images. Therefore, subsequent evaluations can focus on whether these models remain robust when the input distribution shifts from clean digital charts to real camera-captured images.

B. Robustness in Real Camera Images

The main evaluation of this study was conducted on the real camera test set to assess model robustness under practical acquisition conditions. The camera test set contains ammeter chart images captured from screen displays using a mobile phone camera and was not used during training or model selection. This evaluation is important because real camera images may contain perspective distortion, illumination changes, glare, blur, noise, and compression artifacts, which create an acquisition gap between clean digital chart images and field-acquired images.

To evaluate the effect of acquisition augmentation, each architecture was tested under two training conditions: baseline training using only clean digital images and augmented training using the proposed acquisition augmentation strategy. The results are shown in Table VII.

TABLE VII. Camera Test Accuracy: Baseline Training vs Augmented Training

Model	*Baseline Training (%)*	*Augmented Training (%)*	*McNemar p-value*	*Significant?*
SmallCNN	12.0	65.3	6.98×10^{-10}	Yes
MobileNetV3	60.0	77.3	7.20×10^{-3}	Yes
ResNet-50	10.7	89.3	4.32×10^{-14}	Yes
ConvNeXt-Tiny	93.3	93.3	1.00	No

As shown in Table VII, acquisition augmentation substantially improves camera robustness for three of the four evaluated architectures. The largest improvement is observed in ResNet-50, whose accuracy increases from 10.7% under baseline training to 89.3% after acquisition augmentation. SmallCNN also improves from 12.0% to 65.3%, while

MobileNetV3 improves from 60.0% to 77.3%. The McNemar test results show that these improvements are statistically significant, with p-values below 0.05. These results indicate that models trained only on clean digital ammeter chart images are not necessarily robust to real camera images. In particular, SmallCNN and ResNet-50 suffer from severe performance degradation under baseline training, suggesting that the acquisition gap strongly affects their learned representations. After acquisition augmentation is applied, the models become more capable of recognizing current curve patterns despite visual degradations caused by mobile camera acquisition.

ConvNeXt-Tiny achieves the highest camera test accuracy of 93.3%. Interestingly, its accuracy remains the same before and after acquisition augmentation, and the McNemar test indicates that the difference is not statistically significant. This suggests that ConvNeXt-Tiny already has strong generalization capability for camera-captured ammeter chart images, even without additional acquisition augmentation. Therefore, robustness is influenced not only by the augmentation strategy, but also by model architecture and representational capacity.

The confusion matrices in Fig. 3 further support this observation. Models trained without augmentation show more scattered misclassification patterns, especially SmallCNN and ResNet-50. After acquisition augmentation, the predictions become more concentrated along the main diagonal, indicating an increase in correct predictions across most classes. ResNet-50 shows the clearest improvement, changing from a weak baseline model into one of the most robust models on camera images.

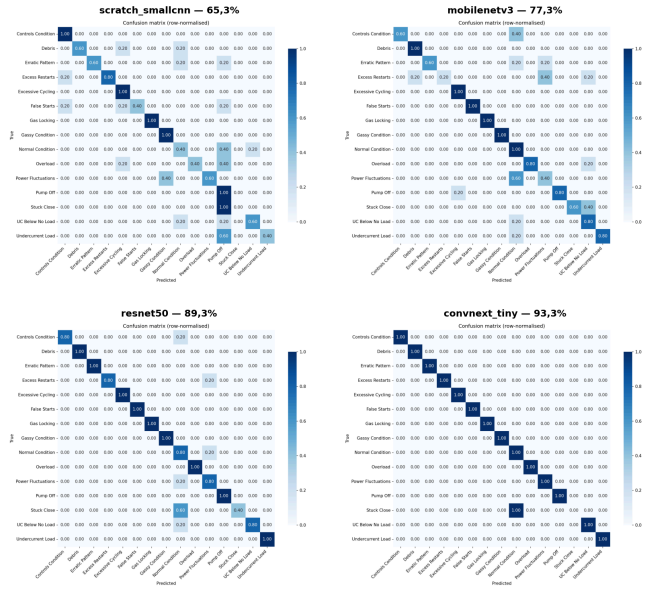

Fig. 3. Confusion matrices of the evaluated models on the real camera test set.

Overall, the camera test results confirm that acquisition augmentation is effective in reducing the impact of the acquisition gap. The results also show that ConvNeXt-Tiny provides the best robustness among the evaluated models, while ResNet-50 benefits the most from the proposed augmentation strategy.

C. *Calibration Analysis*

In addition to classification accuracy, calibration quality is evaluated to determine whether the model confidence is aligned with actual prediction correctness. This aspect is important for a diagnostic system because users may rely on the confidence score to decide whether a prediction can be trusted or should be further verified by an engineer.

Under baseline conditions, the model shows a confidently wrong behavior on camera images. This means that the model often produces high confidence scores even when its predictions are incorrect. Such behavior is undesirable in ESP troubleshooting because it may give users a false sense of reliability. The clean-trained baseline model obtains an Expected Calibration Error (ECE) of 0.66, indicating a large mismatch between predicted confidence and actual accuracy.

After applying acquisition augmentation, calibration quality improves substantially. The augmented ConvNeXt-Tiny model obtains an ECE of 0.071 on the camera test set. This lower ECE indicates that the confidence scores produced by the model are more consistent with the actual correctness of its predictions. In other words, acquisition augmentation not only improves robustness in terms of accuracy, but also helps the model produce more reliable confidence estimates.

The reliability diagram in Fig. 4 illustrates the calibration difference between the baseline condition and the augmented ConvNeXt-Tiny model. The baseline reliability curve shows a large gap between confidence and accuracy, while the augmented model is closer to the ideal calibration line. This result suggests that the proposed augmentation strategy reduces overconfidence under camera acquisition shift.

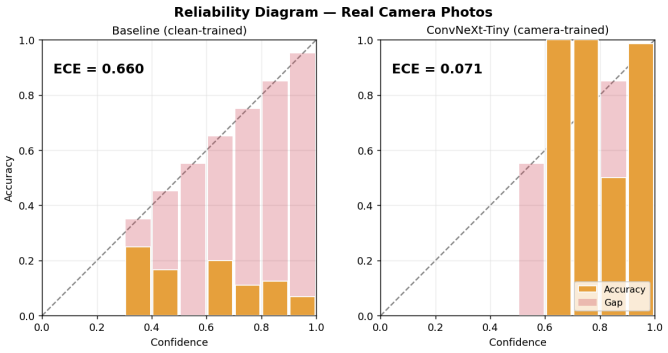

Fig. 4. Reliability diagram on real camera images: baseline condition with ECE of 0.66 and augmented ConvNeXt-Tiny with ECE of 0.071.

This calibration improvement is important for practical deployment. In a field-oriented diagnostic workflow, a well-calibrated model can provide more meaningful confidence information to users. For example, predictions with low confidence can be treated as cases that require further inspection, while high-confidence predictions become more useful as decision support. Therefore, the calibration result strengthens the practical value of the proposed method beyond classification accuracy alone.

D. Latency and Architecture Trade-Off

In addition to classification accuracy, inference latency was measured to evaluate the efficiency of each model. This evaluation is important because the proposed system is intended to support a mobile-based diagnostic workflow, where users may require fast prediction results after submitting an ammeter chart image. The latency and camera accuracy comparison among the evaluated architectures is shown in Table VIII.

TABLE VIII. LATENCY AND CAMERA ACCURACY COMPARISON

Model	*Parameters (M)*	*Latency (ms)*	*Camera Accuracy (%)*
SmallCNN	2.8	2.2	65.3
MobileNetV3	4.2	4.1	77.3
ResNet-50	23.5	8.4	89.3
ConvNeXt-Tiny	28	8.3	93.3

As shown in Table VIII, SmallCNN achieved the lowest inference latency, with 2.2 ms per image. This result is consistent with its smaller parameter count compared with the other evaluated models. However, its camera accuracy was also the lowest among the augmented models, reaching only 65.3%. This indicates that although SmallCNN is computationally efficient, its limited capacity makes it less robust against camera acquisition degradation.

MobileNetV3 achieved a better balance between efficiency and accuracy than SmallCNN. With 4.2 million parameters, it obtained a latency of 4.1 ms and a camera accuracy of 77.3%. This result shows that MobileNetV3 is suitable for scenarios that require relatively fast inference while still maintaining moderate robustness to camera-captured images.

ResNet-50 and ConvNeXt-Tiny achieved substantially higher camera accuracy, but with higher inference latency. ResNet-50 reached 89.3% camera accuracy with 8.4 ms latency, while ConvNeXt-Tiny achieved the best camera accuracy of 93.3% with 8.3 ms latency. Interestingly, although ConvNeXt-Tiny has the largest number of parameters, its latency was slightly lower than ResNet-50 while also producing higher accuracy. This result suggests that architectural design has an important influence on both robustness and computational efficiency.

Overall, the results show a trade-off between model speed and robustness. Lightweight models can provide faster inference, but larger and more modern architectures tend to perform better on real camera images. Based on this evaluation, ConvNeXt-Tiny provides the best balance between accuracy and latency in this study. However, model selection can still be adjusted according to deployment needs. For example, a lightweight model may be preferred when fast response is prioritized, while ConvNeXt-Tiny is more suitable when robustness and diagnostic accuracy are more important.

E. Explainability Analysis

Explainability analysis was conducted to verify whether the model prediction was based on domain-relevant visual regions. In this study, Grad-CAM and LIME were applied to camera-captured ammeter chart images. The purpose of this analysis was to inspect whether the model focused on the current curve pattern, which is the main diagnostic information in an ammeter chart, rather than irrelevant visual components such as labels, borders, grid lines, or background artifacts.

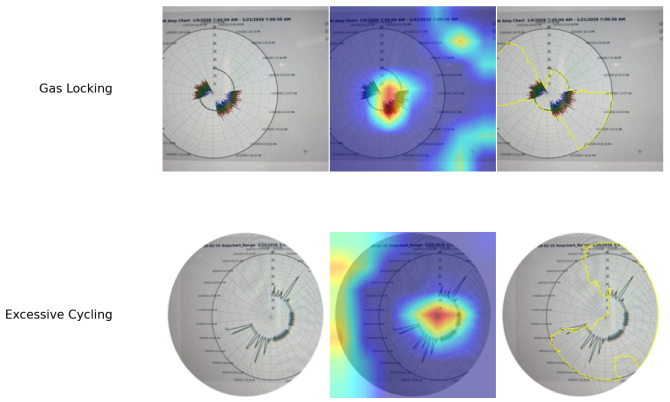

Fig. 5. Grad-CAM and LIME visualizations on camera-captured ammeter chart images, showing model focus on the current curve pattern.

The visualization results are shown in Fig. 5. Grad-CAM produces a heatmap that highlights the regions with strong influence on the predicted class, while LIME highlights superpixel regions that locally support the model prediction. Both methods provide complementary explanations because Grad-CAM is based on internal CNN feature activations, whereas LIME is based on input perturbation.

The results show that Grad-CAM generally highlights the area around the current curve instead of the surrounding chart elements. This indicates that the model decision is not primarily driven by irrelevant background information. In several camera images containing glare, perspective distortion, or illumination variation, the heatmap still remains concentrated around the curve region. This suggests that the model has learned features related to the diagnostic current pattern rather than merely relying on superficial image characteristics.

LIME visualizations also support this finding. The highlighted superpixels generally correspond to regions containing the curve or visually important parts of the chart. These regions contribute to the predicted class and help explain which parts of the image influence the model decision. The consistency between Grad-CAM and LIME strengthens the qualitative evidence that the model focuses on meaningful visual features.

From a practical perspective, the XAI outputs can help engineers inspect the prediction result. If the explanation highlights the current curve, the prediction is more likely to be aligned with domain reasoning. Conversely, if the explanation focuses on irrelevant regions such as text, borders, or background artifacts, the prediction may require further

verification. Therefore, the explainability component provides an additional layer of transparency for the proposed diagnostic system.

F. Error Analysis and Limitations

Although the proposed method achieved strong performance, several errors and limitations were observed. Error analysis was conducted by inspecting misclassified samples on the camera test set, particularly for the best-performing model, ConvNeXt-Tiny.

The main source of error was found in classes with visually similar current patterns. In particular, the Stuck Close class remained difficult to distinguish from Normal Condition. Based on the per-class evaluation, Stuck Close obtained low precision, recall, and F1-score, indicating that the model failed to classify this class correctly in the camera test set. This error occurs because after chart cropping, the absence of current or a relatively flat line in Stuck Close may visually resemble the stable current pattern of Normal Condition. This misclassification is important from an operational perspective because Stuck Close and Normal Condition represent very different ESP states. Normal Condition indicates stable operation, while Stuck Close indicates that the ESP is not operating for a long period. Therefore, errors between these two classes require particular attention before the system is applied in real field scenarios. Further inspection of the misclassified samples shows that the chart cropping process still successfully extracted the chart region. This suggests that the error was not mainly caused by preprocessing failure, but by the visual similarity between the two classes. An example of a misclassified image and its cropped chart region is shown in Fig. 6.

Fig. 6. Example of a misclassified camera image and its cropped chart region.

Several limitations should also be considered. First, the camera test set used in this study is still limited in size. Although the model was evaluated on real camera images, the dataset may not fully represent all possible field variations, such as different camera devices, screen resolutions, lighting conditions, shooting angles, focus quality, and environmental conditions around the well. Second, the acquisition augmentation strategy simulates geometric, photometric, and sensor-related degradation, but it does not explicitly handle extreme occlusion or severe visual obstruction. In real usage, parts of the chart may be covered by strong reflection, shadows, objects, or imperfect cropping. These conditions may remove important curve information and reduce classification performance. Third, the explainability evaluation is still qualitative. Grad-CAM and LIME visualizations were inspected to determine whether the model focuses on the current curve region, but no quantitative faithfulness or stability metric was applied. Future work should include more measurable XAI evaluation to assess whether the explanations are consistently reliable. Fourth, the mobile application currently uses server-side inference. This design reduces the computational burden on the mobile device, but it also depends on network availability and backend reliability. Future development may explore model compression or on-device inference to improve deployment flexibility.

Overall, these limitations indicate that the proposed system is promising but still requires further validation before full field deployment. Future work should expand the real camera dataset, include more diverse acquisition conditions, improve class separation for visually similar patterns, evaluate XAI more quantitatively, and optimize the system for more robust deployment scenarios.

VI. Conclusion

This paper presented an explainable deep learning approach for Electrical Submersible Pump (ESP) troubleshooting using ammeter chart image interpretation. The proposed method addresses the acquisition gap between clean digital ammeter chart images and real camera-captured images through a targeted acquisition augmentation strategy. Four CNN architectures were evaluated, namely SmallCNN, MobileNetV3, ResNet-50, and ConvNeXt-Tiny. The experimental results show that all evaluated models achieved strong performance on clean digital chart images, with ConvNeXt-Tiny obtaining the best clean test performance with 100.00% accuracy and 1.000 macro-F1.

The evaluation on real camera images shows that acquisition augmentation substantially improves model robustness under practical acquisition conditions. The largest improvement was observed in ResNet-50, whose camera test accuracy increased from 10.7% under baseline training to 89.3% after acquisition augmentation. SmallCNN and MobileNetV3 also showed notable improvements, while ConvNeXt-Tiny achieved the best overall camera test accuracy of 93.3%. In addition to improving robustness, the proposed augmentation strategy also improved calibration quality, reducing the Expected Calibration Error from 0.66 under baseline conditions to 0.071 for the augmented ConvNeXt-Tiny model.

The explainability analysis using Grad-CAM and LIME further showed that the model generally focused on the current curve region of the ammeter chart rather than irrelevant visual elements such as chart labels, borders, grid lines, or background artifacts. This indicates that the model predictions are supported by domain-relevant visual patterns, which is important for increasing transparency and supporting engineer verification in ESP troubleshooting. The mobile-based prototype also demonstrates the feasibility of integrating image input, classification, confidence output, and visual explanation into an end-to-end diagnostic workflow.

Future work should expand the real camera dataset to include more diverse devices, screen types, lighting

conditions, shooting angles, and field environments. Further improvements are also needed to distinguish visually similar classes, especially Stuck Close and Normal Condition. In addition, future studies should include quantitative evaluation of explanation quality and explore model compression or on-device inference to improve deployment flexibility in real field conditions.

Acknowledgment

The author would like to express sincere gratitude to Prof. Dr. Ir. Rinaldi Munir, M.T., for his guidance, supervision, and valuable feedback throughout this research. The author also thanks Dr. Fariska Zakhralativa Ruskanda, S.T., M.T., and Ir. Dicky Prima Satya, S.T., M.T., for their constructive evaluation and suggestions. The author gratefully acknowledges PT Pertamina Hulu Energi OSES, especially Mr. Marda Vidrianto, along with the mentors, supervisors, and colleagues who provided domain insights, data support, and practical knowledge related to Electrical Submersible Pump operation and ammeter chart interpretation. Appreciation is also extended to the annotators, academic staff, family, and colleagues who supported the completion of this work.

References

[1] G. Takacs, Electrical Submersible Pumps Manual: Design, Operations, and Maintenance. Cambridge, MA, USA: Gulf Professional Publishing, 2018.

[2] D. Guo, C. S. Raghavendra, K.-T. Yao, M. Harding, A. Anvar, and A. Patel, "Data driven approach to failure prediction for electrical submersible pump systems," in Proc. SPE Western Regional Meeting, Garden Grove, CA, USA, Apr. 2015, doi: 10.2118/174062-MS.

[3] L. Li, C. Hua, and X. Xu, "Condition monitoring and fault diagnosis of electric submersible pump based on wellhead electrical parameters and production parameters," Systems Science & Control Engineering, vol. 6, no. 3, pp. 253–261, 2018, doi: 10.1080/21642583.2018.1548983.

[4] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," Insights into Imaging, vol. 9, no. 4, pp. 611–629, 2018, doi: 10.1007/s13244-018-0639-9.

[5] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," arXiv preprint arXiv:2006.11371, 2020, doi: 10.48550/arXiv.2006.11371.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2019, doi: 10.1007/s11263-019-01228-7.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[8] F. Gong, S. Tong, C. Du, Z. Wan, and S. Qiu, "A fault diagnosis model of an electric submersible pump based on mechanism knowledge," Sensors, vol. 25, no. 8, Art. no. 2444, 2025, doi: 10.3390/s25082444.

[9] Y. Song, Y. Na, K. Kim, T. C. Nguyen, J. Wang, and Y. Kim, "Diagnosis of electrical submersible pump failure using deep learning model with sand-water flow experimental data," Geoenergy Science and Engineering, vol. 243, Art. no. 213279, 2024, doi: 10.1016/j.geoen.2024.213279.

[10] P. Yang, J. Chen, L. Wu, and S. Li, "Fault identification of electric submersible pumps based on unsupervised and multi-source transfer learning integration," Sustainability, vol. 14, no. 16, Art. no. 9870, 2022, doi: 10.3390/su14169870.

[11] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," arXiv preprint arXiv:1506.00327, 2015, doi: 10.48550/arXiv.1506.00327.

[12] M. Valtierra-Rodriguez, J. R. Rivera-Guillen, J. A. Basurto-Hurtado, J. J. De-Santiago-Perez, D. Granados-Lieberman, and J. P. Amezquita-Sanchez, "Convolutional neural network and motor current signature analysis during the transient state for detection of broken rotor bars in induction motors," Sensors, vol. 20, no. 13, Art. no. 3721, 2020, doi: 10.3390/s20133721.

[13] H. Zhang, C. Liu, Z. Zhang, Y. Xing, X. Liu, R. Dong, Y. He, L. Xia, and F. Liu, "Recurrence plot-based approach for cardiac arrhythmia classification using Inception-ResNet-v2," Frontiers in Physiology, vol. 12, Art. no. 648950, 2021, doi: 10.3389/fphys.2021.648950.

[14] M. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Explainable AI approaches in deep learning: Advancements, applications and challenges," Computers and Electrical Engineering, vol. 117, Art. no. 109246, 2024, doi: 10.1016/j.compeleceng.2024.109246.

[15] S. Moosavi, M. Farajzadeh-Zanjani, R. Razavi-Far, V. Palade, and M. Saif, "Explainable AI in manufacturing and industrial cyber–physical systems: A survey," Electronics, vol. 13, no. 17, Art. no. 3497, 2024, doi: 10.3390/electronics13173497.

[16] E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," Applied Sciences, vol. 13, no. 4, Art. no. 2038, 2023, doi: 10.3390/app13042038.

[17] A. Howard, M. Sandler, G. Chu, et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016.

[19] Z. Liu, H. Mao, C.-Y. Wu, et al., "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2022.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, Art. no. 60, 2019.

[23] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, 2017.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[25] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. 29th AAAI Conf. Artificial Intelligence, 2015.

[26] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.