EPJ Photovoltaics **16**, 23 (2025)

https://doi.org/10.1051/epjpv/2025010

Available online at:
www.epj-pv.org

Special Issue on 'EU PVSEC 2024: State of the Art and Developments in Photovoltaics', edited by Robert Kenny and Gabriele Eder

ORIGINAL ARTICLE OPEN ACCESS

Redefining failure detection in PV Systems: a comparative study of GPT-4o and ResNet's computer vision in aerial infrared imagery analysis

Sandra Gallmetzer[1,*], Mousa Sondoqah[1], Evelyn Turri[1,2], Lukas Koester[1], Atse Louwen[1], and David Moser[1]

[1] Institute for Renewable Energy, Eurac Research, Viale Druso 1, 39100 Bolzano, Italy
[2] Università degli Studi di Modena e Reggio Emilia, Dipartimento di Ingegneria "Enzo Ferrari", Via P. Vivarelli 10, 41125 Modena MO, Italy

Received: 9 December 2024 / Accepted: 7 March 2025

Abstract. The rapid growth of the solar photovoltaic industry underlines the importance of effective operation and maintenance strategies, particularly for large-scale systems. Aerial infrared thermography has become an essential tool for detecting anomalies in photovoltaic modules due to its cost-effectiveness and scalability. Continuous monitoring through advanced fault detection and classification methods can maintain optimal system performance and extend the life of PV modules. This study investigates the application of advanced artificial intelligence methods for fault detection and classification comparing the performance of GPT-4o, a multimodal large language model, and ResNet, a convolutional neural network renowned for image classification tasks. Our research evaluates the effectiveness of both models using infrared images, focusing on binary defect detection and multiclass classification. ResNet demonstrated advantages in terms of computational efficiency and ease of implementation. Conversely, GPT-4o offered superior adaptability and interpretability, effectively analysing multimodal data to identify and explain subtle anomalies in thermal imagery. However, its higher computational requirements limit its feasibility in resource-limited settings. The results highlight the complementary strengths of these models and provide valuable insights into their role in advancing automated fault diagnosis in photovoltaic systems.

Keywords: Inspection / performance / failure detection and classification / photovoltaics / multimodal large language models / prompt engineering

1 Introduction

The solar photovoltaic (PV) industry has experienced significant and rapid growth in recent years, driven by the need for renewable energy sources and global efforts to reduce carbon emissions. As PV systems become more prevalent, ensuring constant and reliable energy output requires efficient operation and maintenance (O&M). Effective O&M strategies are crucial for maintaining high availability and reliability, which directly impact the financial returns of PV plants.

As the PV industry matures, the expected operational lifetime of PV modules increases [1]. With aging comes an increased occurrence of faults and performance degradation [2], highlighting the need for effective detection methods.

Many recent studies [3,4] highlight the growing role of PV module imaging in PV system maintenance as a cost-effective, less time-consuming alternative and valuable complement to traditional electrical performance tests. Image-based inspection methods are automatable, scalable, and thus particularly beneficial for large-scale PV plants [5]. In particular, in the field of PV plant monitoring, imaging techniques have significant potential to provide autonomous solutions [6]. Fast and accurate detection, localization, and classification of anomalies in solar modules and other PV plant components not only reduces downtime [7] but also lowers the associated costs of repair, ultimately improving the overall operational efficiency and costs of PV systems [8].

The O&M best practice guidelines from the PVPS Task Force [2], SolarPower Europe [9], and NREL [10], recommend infrared thermography (IRT) as a key image-based inspection method, collectively highlighting its critical role in efficient PV system diagnostics and maintenance.

* e-mail: sandra.gallmetzer@eurac.edu

Infrared thermography has been an invaluable tool for identifying hotspots and diagnosing defects in PV modules, making it a widely adopted technique for field testing and PV device characterization [11]. The shape and location of the hotspot on a PV module can give an indication of its root cause, enabling in some cases defect classification. The IEC standard 62446-3 [12] describes these infrared patterns and outlines requirements for testing, documentation, and maintenance through outdoor IRT, ensuring consistent and accurate detection and classification of module defects. In recent years, IRT inspection advanced to better solutions by applying automated and robotic solutions. Aerial IRT has recently been applied using unmanned aerial vehicles (UAV), offering the advantages of faster application speeds and reduced costs compared to ground-level inspections [13], making it a powerful tool for large-scale PV plant maintenance.

Current analysis of IRT images for PV systems focuses mainly on hot spot detection, with specific defect classification still underdeveloped. Manual approaches, such as temperature line profiles and histogram-based statistical analysis of regions of interest (ROI), are commonly used for the qualitative assessment of indoor and outdoor measurements [14]. Tools such as FLIR Thermal Studio, IRBIS and IRimage allow manual or automated analysis, while drone-specific platforms such as DroneDeploy and DJI Terra enhance aerial inspections. Advanced approaches include the use of Python libraries for custom pipelines or the MATLAB statistical [15] and machine learning toolbox [16]. Signal processing algorithms are used to reduce noise in thermal images [14].

Machine learning methods, such as hybrid feature-based support vector machines (SVMs) for hot spot detection [16] and convolutional neural network (CNN)-based models for automatic detection and classification [5,17], are emerging but still need to be developed for full defect classification.

However, despite advancements in technology, the detection and classification of potential failures in PV modules remains a time-consuming and resource-intensive task, particularly for utility-scale PV plants. While robotic solutions are becoming more and more trendy for image captioning [18], there is a pressing need to automate image processing itself due to the growing volume of image datasets and the high impact of personnel costs required for manual analysis [8]. Artificial intelligence (AI)-driven solutions might be a possibility for further automation in the image processing task, which is being addressed in this study.

In this work, we present a comparative study of GPT-4o and ResNet's Computer Vision in the context of aerial infrared thermography analysis for PV systems. GPT-4o, a multimodal large language model (MLLM), offers advanced capabilities for image analysis and has already shown promising capabilities in various non-PV-related applications [19,20]. On the other hand, ResNet, a well-established deep learning model known for its performance in image classification tasks [21], recently applied also in the health sector for disease classification [22], serves as a benchmark for evaluating the performance of GPT-4o.

The structure of this study is organized as follows: Section 2 provides a detailed literature review on the current state-of-the-art techniques used in PV system fault detection and fault classification. Section 3 outlines the methodology applied in this study, focusing on the comparative performance of GPT-4o and ResNet. In Section 4, we present and analyze the results, followed by a discussion of the findings and some best practices for the use of multimodal large language models within the field of defect analysis for PV. Finally, conclusions are drawn in Section 5, highlighting the implications of this study for future PV system monitoring and maintenance.

2 Literature review

This work addresses defect detection and classification, making it essential to distinguish the terms *detection* and *classification* clearly from each other and provide the definition which is used in this study.

Defect *detection* is the process of identifying anomalies or deviations from the normal operation of a system or component. According to Köntges et al. [23] detection in the context of photovoltaic systems focuses on locating areas on the PV module presenting irregularities, such as hot spots, broken cells, or electrical faults, without necessarily specifying the type or cause of the anomaly. The primary objective of detection is to identify potential problems for further investigation, enabling operators to take preventive or corrective action to maintain system efficiency and avoid more serious failures [5]. Therefore, the term *detection* is used to define the presence of an anomaly within a PV panel. In contrast, the term *classification* describes the further analysis of a present anomaly to identify the type of anomaly [24]. Through classification of the detected defect, we are providing a diagnosis that is useful to know in order to be able to apply the best solution for fixing the issue [8].

Defects that cause performance losses and can be identified by IRT have been listed by Koester et al. [25] in a matrix. This matrix shows that hot spots can be caused by bypass diode short circuits, modules in open-circuit conditions appearing offline, potential-induced degradation (PID), major glass breakage, or fractured solar cells. The corresponding infrared patterns are described in the IEC standard [12].

In a study by Fernandez et al. [6] the authors propose a fully automated approach for the detection, classification, and geolocalization of faults visible in thermographic images of large-scale PV plants. Their work emphasizes the importance of robust image processing techniques for effective fault detection. Although the proposed method achieved significant improvements in operational efficiency, the paper notes the complexity of manual image analysis and the need for advanced algorithms to enhance the automation of failure detection.

Herraiz et al. [26] proposed a CNN-based approach for detecting faults in thermographic images captured by UAVs. Their study employed two region-based CNNs to identify photovoltaic modules exhibiting hot spots, a common indicator of defects. This method demonstrates

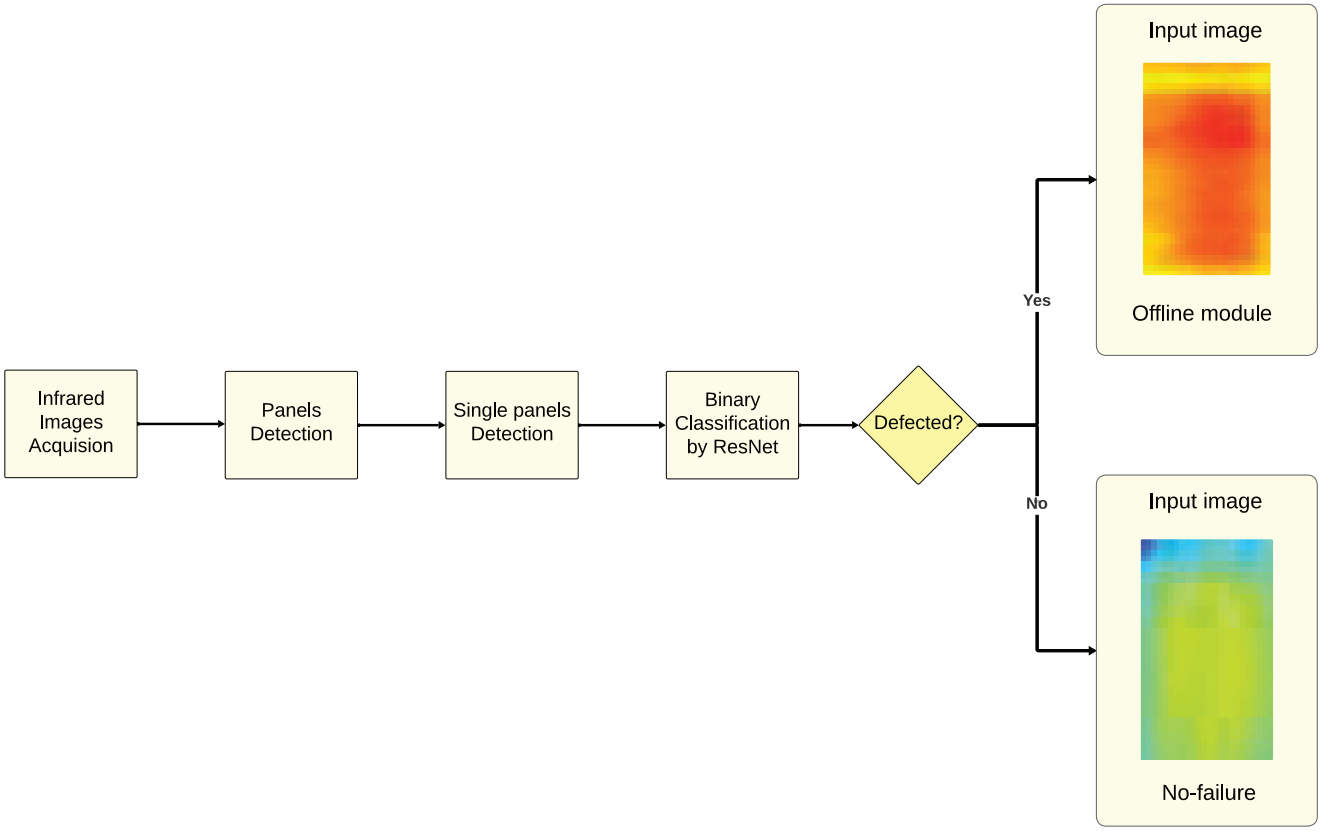

Fig. 1. Data processing flow for failure detection and classification using ResNet and GPT-4o.

the effectiveness of CNN architectures in automatically detecting faults in thermal imagery. Classification of defects, instead, was not performed in their work.

Building upon previous research [27,28], Alves et al. [24] advanced beyond fault detection by incorporating the classification of various types of defects. This approach enhances the capability of CNNs to not only detect but also distinguish between different fault categories, offering a more comprehensive analysis of thermographic data.

This research provides a solid foundation for exploring and comparing advanced machine-learning models applied to the world of photovoltaic diagnostics. While CNN architectures like ResNet have proven effective in detecting and classifying defects visible in IRT images, recent developments in MLLMs offer the possibility to integrate and process multiple data types, like images, text, and audio. Wu et al. [20] and Zhang et al. [29] describe the challenges and latest advancements of such models including VisualChatGPT, MiniGPT-4, and LLaVA which were developed in 2023 and are capable of performing visual understanding tasks.

From a computational perspective, the requirements of a large language model differ significantly from computer vision models. The latter relies on convolutional operations optimized for parallel processing, making them relatively efficient for deployment across various hardware, as analyzed in studies like Desislavov et al. [30]. In contrast, LLMs, such as GPT-based architectures, are transformer-based models that require extensive memory and compute resources, particularly for autoregressive token generation. Studies such as Wu et al. [20] have explored accelerators to optimize transformer efficiency, while other analyses [31] compare pre-trained models in terms of their computational trade-offs. The operational costs of LLMs tend to be significantly higher due to their vast parameter sizes and sequential processing requirements, making considerations of energy efficiency and hardware specialization crucial in model selection. The final choice between these model types depends on factors such as task specificity, computational constraints, and deployment scalability.

3 Methodology

Building on this knowledge, this paper examines and compares the two models, ResNet [32] and GPT-4o [33], in the analysis of infrared aerial images for defect detection and classification in PV systems. Before deep diving into each of them, it is important to clarify the two tasks in the deep learning field as well. Failure detection is the task of detecting if there is a failure or not in the infrared image, in deep learning field this task is equivalent to a binary classification of the images, where the goal is to assign to each a class between {no-failure, failure}, where failure is the class describing an image with any type of failure, while no-failure means that the infrared image does not have any of them. The corresponding data processing flow is shown in Figure 1. Mathematically, given the dataset

$$\mathscr{D} = \{x_i, y_i\}_{i=1}^{N} \tag{1}$$

where x_i is the image i in the dataset, where $x_i \in \mathcal{X}_{binary}$, y_i is the label ground truth i for the sample x_i, where $y_i \in \mathcal{Y}_{binary}$ and $\mathcal{Y}_{binary} = \{0, 1\}$, where 0 is the class for no-failure and 1 is the class for failure, N is the number of samples in the binary dataset.

The task can be formalized as learning the function f_{binary}:

$$f_{binary} : \mathcal{X}_{binary} \rightarrow \{0, 1\}. \tag{2}$$

Failure classification, instead, is the task of multiclass classification in deep learning, where each class is a type of failure and the model assigns a class to each of the images, based on what kind of failure is present. Mathematically we can formalize the task as learning the function f_{multi}:

$$f_{multi} : \mathcal{X}_{multi} \rightarrow \{0, ..., K-1\} \tag{3}$$

where K is the number of classes for the multi-class classification task and $\mathcal{X}_{multi}$ is the dataset containing the images with a failure.

For the detection task, accuracy is used as a performance metric. It measures the proportion of correctly identified defective and non-defective images over the total number of predictions and is mathematically expressed as:

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

where: TP is the number of **true positives** (correctly identified defective images) TN is the number of **true negatives** (correctly identified non-defective images) FP is the number of **false positives** (non-defective images wrongly classified as defective) FN is the number of **false negatives** (defective images wrongly classified as non-defective).

For the classification task, where an image must be assigned to one of multiple defect categories, accuracy is computed as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}. \tag{5}$$

Since this is a multi-class classification problem, accuracy measures the proportion of correctly classified images among all predictions. Unlike detection accuracy, this metric does not include true negatives, as every image must be assigned a defect class. If the classification task involved multiple defect labels per image, a more appropriate metric would be subset accuracy:

$$\text{Subset Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{\mathbf{y}}_i = \mathbf{y}_i) \tag{6}$$

where N is the total number of instances in the dataset, y_i is the ground-truth binary label vector for instance i, $\mathbf{y}_i$ is the predicted binary label vector for instance i, $-\mathbb{1}(\cdot)$ is the indicator function, which returns 1 if the prediction exactly matches the ground truth and 0 otherwise.

Both tasks have been evaluated using two different models, ResNet and GPT-4o, as described in the following subsections.

3.1 Application of ResNet

ResNet is a deep CNN introduced by He et al. [32] during the Imagenet competition[1]. Their winning innovation is the addition of the residual blocks to the network, which allows the model to learn residual mappings, so instead of learning the desired output, they learn the difference between the input and the output, which is more efficient than the traditional method. The CNNs are networks able to learn patterns in the input image, and for this reason, are excellent for image classification tasks. The main parts of the ResNet are convolutional layers, residual blocks, fully connected layers, and an output layer. The convolutional layers apply a series of filters to the input image to extract relevant spatial features, these filters or kernels slide into the image, by highlighting patterns such as edges or textures. As previously written, the key feature of the ResNet is the residual block, which allows the output of the convolution to skip some layers thanks to the skip connection or shortcut and ensures that the output is not degraded. Instead, fully connected layers help in learning non-linear combinations of features after convolutions and residual blocks. As the final part, there is the output layer, which in this case is a softmax and output the features into class probabilities.

Giving a mathematical formulation, in the end, the ResNet model outputs class probability for each sample, so $\hat{y}_i = \mathrm{P}(y = i|x_i)$ where $i \in \{0,\ C-1\}$ and C is the number of classes, in the binary classification case $C = 2$ while in the multi-class classification case $C = K$. In particular, suing the softmax in the output layer the class probability formula is entirely written as:

$$\hat{y}_i = \mathrm{P}(y = i|x_i) = \frac{e^{w_i^T x + b_i}}{\sum_{j=1}^{C} e^{w_j^T x + b_j}} i \in \{0, ..., C-1\} \tag{7}$$

where: w_i R^d is the weight vector associated with class i, $b_i \in$ R is the bias term for class i, $e^{w_i^T x + b_i}$ is the exponential of the score (logit) for class i, and $\sum_{j=1}^{C} e^{w_j^T x + b_j}$ is the sum of the exponentials of the logits for all classes, ensuring that the output is normalized into a probability distribution.

After computing the class probabilities, which is the output of the ResNet, the class assigned is the one with the highest probabilities, so:

$$\hat{y}_{class} = \underset{i}{\text{argmax}}\, \hat{y}_i. \tag{8}$$

To implement this formulation, a transfer learning technique has been used, meaning that the backbone of the ResNet was already pre-trained using the weights coming from the training of the network over the ImageNet dataset [34], while only the output layer of the model has been trained over the Infrared images dataset for the specific

[1] https://www.image-net.org/index.php

task of binary and multi-class classification. In particular, in the work, three different types of ResNet have been trained, ResNet18, ResNet50, and ResNet101. The main difference between them is the depth of the network, therefore how many layers have been used for the model. This difference implies the diverse capacity of the networks and the capability to learn complex features. A bigger model, such as ResNet101, is not always better, because it depends as well on the complexity of the task. Indeed, if the goal to achieve is easy, a smaller network might have better generalization capabilities than a bigger one. For this reason, all three models have been trained for the different tasks, in order to have a better analysis of the performances.

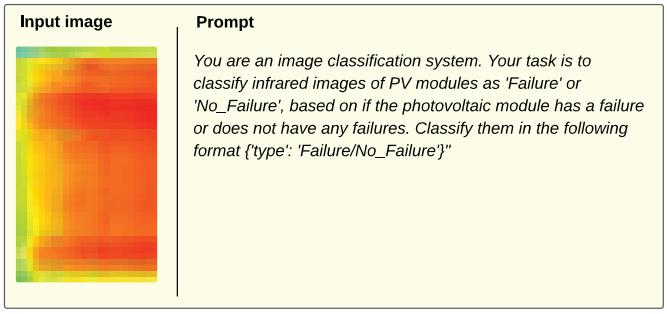

Fig. 2. GPT-4o prompt engineering for failure detection.

3.2 Application of GPT-4o

MLLMs incorporate a variety of data types that go beyond the limitations of purely text-based models and enable the processing of different data forms. GPT-4o is a prime example of such an MLLM. It is capable of processing input in the form of both images and text, showcasing performance on par with humans in numerous benchmark tests [20]. To exploit the capabilities of the MLLM GPT-4o in the context of failure detection and classification in infrared images, we first investigate the performance of the model without additional guidance to observe its behavior in an unsupervised context. In the second phase, we apply prompt engineering and provide detailed descriptions of infrared patterns based on established standards and relevant literature [12,23,25]. To conduct this study, we test GPT-4o through the application programming interface (API). For the input format of the GPT-4o model, prompt text is provided alongside single images at a time. However, for reproducibility, batching can be used to input multiple images as a single file. These images are processed along with a system prompt, which can be cached to reduce inference costs. The GPT-4o model outputs structured data, allowing results to be formatted in JSON using "response_format": "json" as a given parameter. This enables the model to return a final label for each input image (e.g., "type": "Hot-Spot"), simplifying post-inference analysis and evaluation.

3.2.1 Prompt engineering

Various prompting strategies can be employed to utilize multi-modal large language models for tasks involving diverse data types. Instructional prompting facilitates zero-shot learning, enabling models to perform tasks based solely on explicit instructions, without task-specific examples. To improve task generalization, few-shot prompting can be utilized, providing the model with a limited set of labeled image examples, allowing it to infer patterns and apply them to new, unseen data [35]. For more complex reasoning tasks, Chain of Thought (CoT) prompting [36] guides the model through a step-by-step logical process, improving its ability to handle intricate problems. Visual grounding techniques can be used to direct the model's attention to specific regions of visual inputs, aiding in the interpretation of complex scenes or data [37]. To further enhance task-specific performance, fine-tuning the model with domain-relevant datasets can improve its capacity to recognize specific patterns and features [38]. These approaches collectively enable MLLMs to effectively process and analyze a wide range of data modalities, from textual information to visual content, and perform diverse tasks with increasing accuracy and sophistication.

In this study, we employed GPT-4o for both detection and classification tasks using various approaches. For the detection task, we implemented a zero-shot prompting strategy to determine the presence or absence of failures in thermal images as shown in Figure 2. This binary classification task was intentionally designed without providing explicit information about the two classes, allowing the model to make independent determinations. The quality of results from Large Language Models (LLMs) is heavily dependent on prompt engineering. As demonstrated, careful prompt refinement is essential for obtaining accurate results from these models [39]. Moreover, well-crafted prompts can significantly enhance both model performance and contextual understanding [40].

After careful iteration, we developed the following prompt for the detection task:

"You are an image classification system. Your task is to classify infrared images of PV modules as 'Failure' or 'No_Failure', based on if the photovoltaic module has a failure or does not have any failures. Classify them in the following format 'type': 'Failure/No_Failure'"

For the defect classification task, we implemented two distinct prompting approaches. The first was a zero-shot approach, similar to our detection task, where we prompted the model to classify the image into specific defect categories without providing additional contextual information. The final prompt used for this classification task is shown in Figure 3.

The second approach also utilized zero-shot prompting but incorporated detailed definitions and visual characteristics for each failure type. This instructional prompting approach, while still maintaining zero-shot learning principles, provided the MLLM with explicit descriptions of the visual properties specific to our testing dataset. This method enables the model to perform classification tasks based on detailed instructions alone, without requiring task-specific examples. For example, a description of the hotspot failure was as follows: "Hot-Spot is a PV module

Table 1. Applied descriptions of infrared patterns for the PV module defects of interest for this study.

Defect	Pattern description
Cell	A PV module with a single clearly defined square cell appearing hotter than the surrounding cells.
Cell-Multi	A PV module with several distinct square cells randomly distributed over the module that are noticeably warmer than others.
Cracking	A PV module with distinct bright spots or lines indicating localized heat generation, potentially of irregular shapes.
Hot-Spot	A PV module with a single small point-like, localized, intense heating on one part of the module, with the surrounding region gradually transitioning back to a normal temperature.
Hot-Spot-Multi	A PV module with multiple hotspots spread across different areas of the module. Each hotspot appears much warmer than the surrounding areas.
Shadowing	A PV module with uneven heating patterns that appear random and non-uniform, with a gradual transition between cooler and warmer areas across the module.
Diode	A PV module with a single bypass diode failure, which has the pattern of a single row or column appearing warmer than the rest of the module.
Diode-Multi	A PV module with several large areas or strings of cells appearing warm. These inactive regions may span several adjacent rows or sections of the module.
Vegetation	A PV module with one or more areas, typically on the borders of the module, appearing cooler, with the surrounding areas appearing warmer.
Soiling	A PV module with a uniform cooling effect across the affected areas, typically on the borders of the module.
Offline-Module	A PV module with a uniform pattern across the entire module, with no noticeable hotspots or irregularities.

with a single small point-like, localized, intense heating on one part of the module with the surrounding region gradually transitioning back to a normal temperature". Table 1 summarizes all IR pattern descriptions used for this study.

3.3 Dataset

This experiment utilizes the publicly available InfraredSolarModules Dataset collected by Millendorf et al., [41] which consists of 20,000 infrared images of solar modules, each measuring 24 by 40 pixels. The dataset is divided into twelve classes, including eleven distinct anomaly classes and a "No-Anomaly" class for modules without defects. Of the total images, 10,000 exhibitarious faults or anomalies, while the remaining 10,000 show no defects [41].

The images were captured using midwave or longwave infrared (3–13.5 μM) thermography cameras mounted on UAVs and piloted aircraft [41]. The resolution of the images varies from 3.0 to 15.0 *cm/pixel.*, nd each image was cropped to focus on an individual solar modules [41]. This dataset provides real world, labeled data with unique examples of solar module anomalies. In addition to infrared imaging, visible spectrum images were used during classification to improve accuracy.

Figure 4 presents the unbalanced distribution of the defect classes for this dataset, which although reflects the real-world occurrence of failures detectable through IRT [25,42].

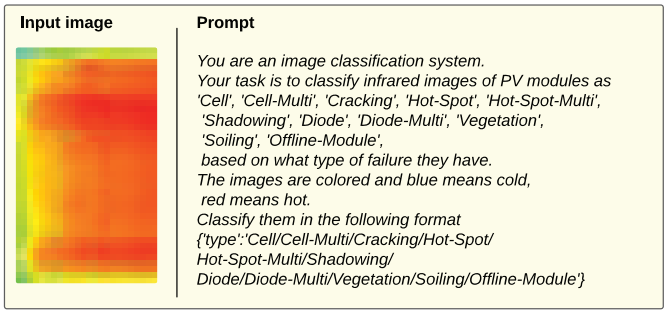

Fig. 3. GPT-4o prompt engineering for failure classification.

The infrared images in the dataset used for this study are originally presented in grayscale, with some sample images shown in Figure 5. Upon reviewing these samples, it becomes evident that certain defect classes, such as diode and diode-multi, exhibit similar patterns, making them difficult to distinguish. Conversely, other classes, like vegetation, shadowing and soiling, display significant intra-class variation, further complicating the classification process explained more in detail in Section 4.2. To address these challenges, we applied the Jet colormap to the grayscale images. This colormap improves the visual representation by dividing the spectrum into multiple colors, making subtle differences and fine details more perceptible. This transformation is the key reason behind our different accuracy results between color and grayscale

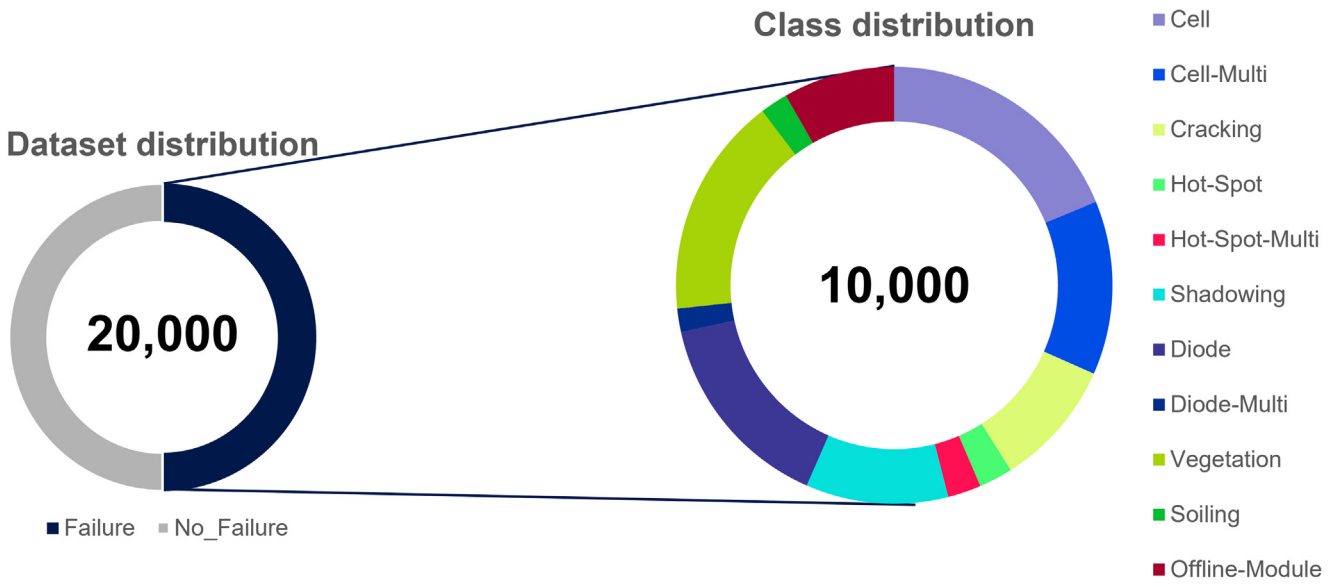

Fig. 4. Class distribution of the InfraredSolarModules dataset [41].

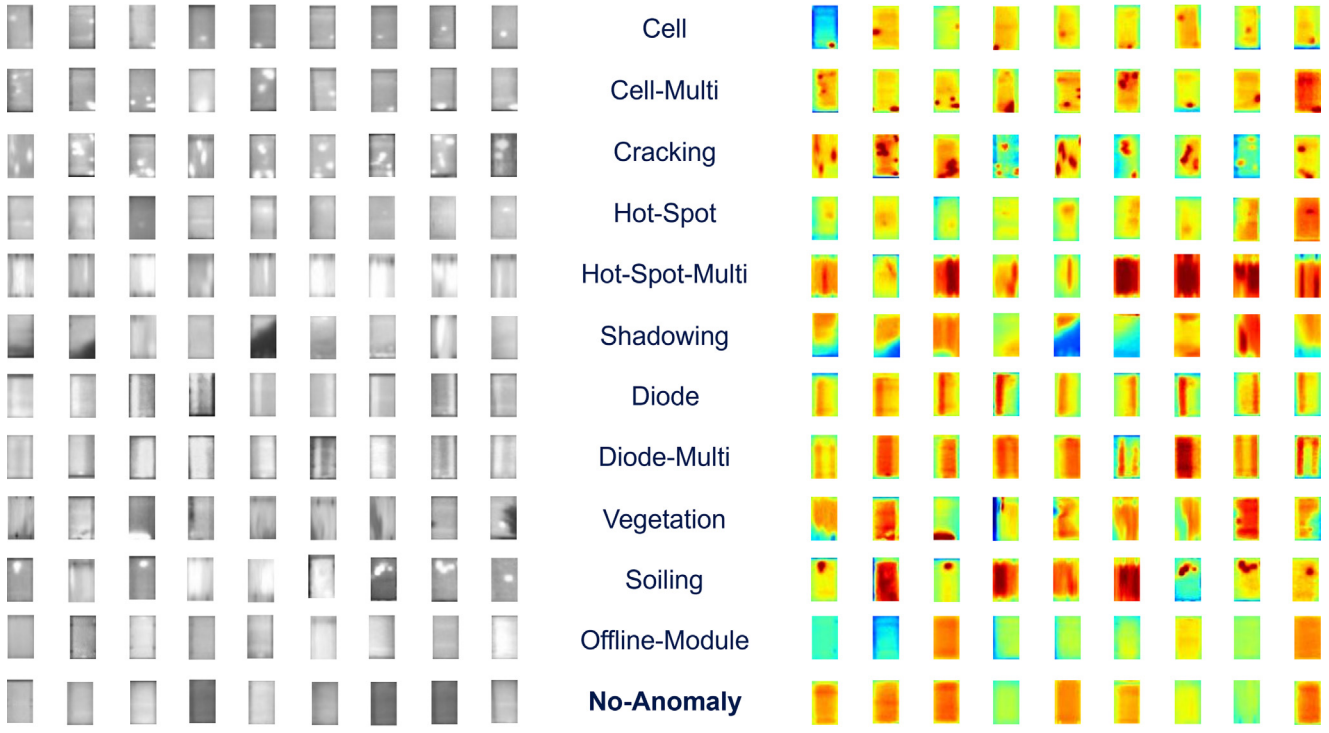

Fig. 5. Sample images from the InfraredSolarModules dataset [41], divided into 12 anomaly classes. The grey-scale images are shown on the left, and the JetColor map transformations are displayed on the right.

processing approaches. When processing grayscale images, the model extracts features from a single intensity channel, while colored images provide three channels of distinct patterns and details. The accuracy differences vary across different classes of failures – while color features enhanced the detection of certain defect types through additional properties like hue and saturation, the simpler grayscale representation proved more effective for other defects by focusing only on essential structural patterns without the potential distraction of color variations.

The difference in accuracy between color and grayscale image processing reflects their fundamental differences in feature extraction. Grayscale images, with their single channel, rely on intensity information. In contrast, color

Table 2. Results for detection across different models.

Accuracy – Detection					
	ResNet			GPT-4o	
	ResNet18	ResNet50	ResNet100	Zero-shot prompting	Instructional prompting
Grey-scale	0.80	0.79	0.81	0.72	0.72
Colored		0.85	0.84	0.73	0.69

Table 3. Results for classification of all eleven failure classes.

Accuracy – Classification for 11 failure classes					
	ResNet			GPT-4o	
	ResNet18	ResNet50	ResNet100	Zero-shot prompting	Instructional prompting
Grey-scale	0.48	0.61	0.60	0.10	0.16
Colored		0.66	0.62	0.12	0.24

images provide three channels, each capturing distinct patterns and details. While more channels can offer richer features for classification through properties like hue and saturation, fewer channels may sometimes prove beneficial by forcing the model to focus on essential patterns rather than potentially misleading color variations. The effectiveness of each approach varies depending on the specific task at hand.

4 Results

The testing of the two models, ResNet and GPT-4o, is divided into two phases: detection in the first phase and classification in the second one.

4.1 Results for detection

In the first phase of testing, which focuses on the detection, the goal is to determine whether a defect is present in the PV module. Table 2 shows the accuracy results for the detection task across the grey-scale and colored datasets. The performance is evaluated for both ResNet and GPT-4o, with different configurations: ResNet18, ResNet50, ResNet100, GPT-4o with zero-shot prompting and GPT-4o applying instructional prompting. Among the three computer vision models, ResNet100 demonstrates the highest accuracy for the grey-scale dataset, with an accuracy of 81%, while ResNet50 and ResNet18 show slightly lower accuracies of 79% and 80%, respectively. Comparatively, GPT-4o's performance is slightly lower than the ResNet models. The accuracy for zero-shot prompting is 72% for both grey-scale and colored datasets, while through the application of instructional prompting, GPT-4o performs similarly with 72% for grey-scale and 69% for colored.

Overall, the results highlight the strengths of the ResNet models in detecting anomalies, while GPT-4o's performance is relatively consistent, but lags behind. It may require further fine-tuning for better performance. In the following section, we will analyze the performance of the different models in the classification of defects visible on an IRT image.

4.2 Results for classification

For the classification task of defects, the testing is divided into two cases. The first case considers all eleven failure classes and treats them as individual classes. Table 3 summarizes the results of this test. Among the ResNet models, ResNet50 shows the best performance in classifying IR defects for both grey-scale and colored datasets, with an accuracy of 61% and 66%, respectively. ResNet18, instead, reaches only 48% of accuracy in the grey-scale leading to the conclusion that a bigger model is needed. Finally, ResNet100, with an accuracy of 60% and 62% for the grey-scale and the colored datasets, performs also worse than ResNet50. These results indicate that increasing the model depth from ResNet50 to ResNet100 does not lead to a further improvement in this specific classification task. For GPT-4o, the performance is considerably lower compared to the ResNet models. The accuracy of GPT-4o for zero-shot prompting is 10% for the grey-scale dataset, indicating that GPT-4o has limited effectiveness in classifying the defects of the underlying dataset. With instructional prompting, GPT-4o's performance improves slightly, but it still falls short of the ResNet models, with an accuracy of 16% for the grey-scale dataset. On the colored dataset, the results are similarly low, achieving 12% accuracy with the zero-shot and 24% with the instructional prompting.

Table 4. Results for the reduced dataset classifying only five failure classes.

Accuracy – Classification for 5 failure classes					
	ResNet		GPT-4o		
	ResNet18	ResNet50	ResNet100	Zero-shot prompting	Instructional prompting
Grey-scale	0.54	0.86	0.87	0.33	0.63
Colored	0.60	0.86	0.87	0.36	0.71

In the second classification case, the dataset is reduced to only five defect classes. Classes showing high variation in the infrared patterns within a single failure category, such as shadowing, vegetation, soiling, and cracking, are eliminated from the pool of images. These defect classes do not present a clear IR pattern repeatable over the data in the same class therefore it is hard even for classical computer vision models to learn them, as it was confirmed by Alves et al. [24] through the provision of the confusion matrices. The classes we excluded can in most cases not be detected by only thermal images but would benefit from visual image analysis or even luminescence image analysis to get a deeper insight into the failure pattern structure. The defect classes selected for this second test are therefore: offline module, hotspot, multi-hotspot, diode, and multi-diode. The models were retrained and retested accordingly with the reduced sample of images. Table 4 presents the main results for each model:

ResNet100 achieves the highest performance, with a testing accuracy of 87% on both, the grey-scale and the colored dataset, closely followed by ResNet50 with an accuracy of 86%. ResNet18, instead, falls short of the other two models, achieving accuracies of 54% for the grey-scale and 60% for the colored dataset.

Similar to ResNet, GPT-4o also shows improved performance with the reduced dataset. The instructional prompting achieves accuracies of 63% for the grey-scale and 71% for the colored dataset while for zero-shot prompting, the model attains only 33% and 36%, respectively.

GPT-4o demonstrates high reliability in classification tasks, with minimal variability across multiple executions. Our evaluation across ten runs reveals a standard deviation of 0.035 in accuracy, indicating that the model's predictions remain largely stable. Additionally, the per-sample variance is near 0.0, reinforcing consistency in individual classifications. Figure 6 visually represents the accuracy distribution across ten runs, further supporting this finding. While these results suggest GPT-4o can be reliably applied in many scenarios, minor fluctuations may arise due to input sensitivity or inherent stochastic behavior in transformer-based models. For critical applications, such as medical or safety-related decision-making, incorporating confidence thresholds or ensemble methods could help mitigate any residual variability.

Overall, the results demonstrate the superior performance of ResNet models across both detection and classification tasks, with ResNet100 consistently achieving the highest accuracies. GPT-4o shows promise, especially when instructional prompting is applied, but its performance lags behind ResNet, particularly for fine-grained classification. While detection tasks yield high accuracies overall, classification poses greater challenges, especially with many different defect classes showing similar infrared patterns. Simplifying the dataset to fewer defect classes significantly improves performance for both approaches, highlighting the importance of dataset quality optimization and tailored guidance in addressing the complexities of photovoltaic defect analysis with GPT-4o.

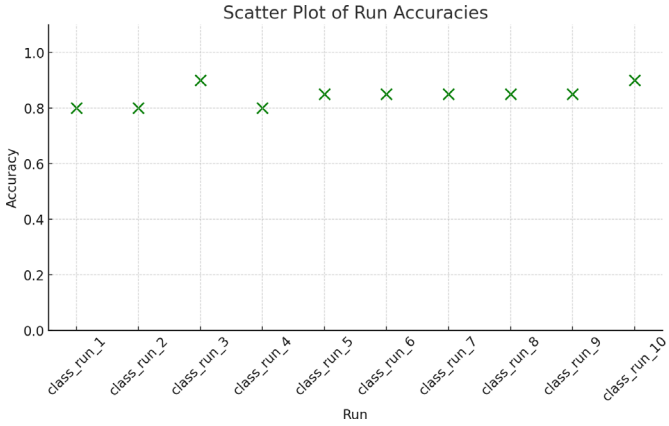

Fig. 6. The different values of classification accuracy for GPT-4o through the different runs.

4.3 Best practice for defect detection with GPT-4o

The application of MLLM for failure detection and classification in PV systems revealed several significant insights. The initial evaluation was conducted on a low-resolution dataset with images of 24×40 pixels, which presented challenges for GPT-4o's pattern recognition capabilities, as the model cannot be specifically fine-tuned for such constrained input dimensions. The issue with capturing low-resolution images is that their feature extraction will be limited and suboptimal. The vision encoder in GPT-4o likely performs best with sufficiently high-resolution images, as smaller images may produce poor embeddings. This can result in objects appearing blurred or pixelated, leading to inaccurate or vague responses. A notable advantage of the MLLM approach is its efficiency in handling raw image data. Unlike

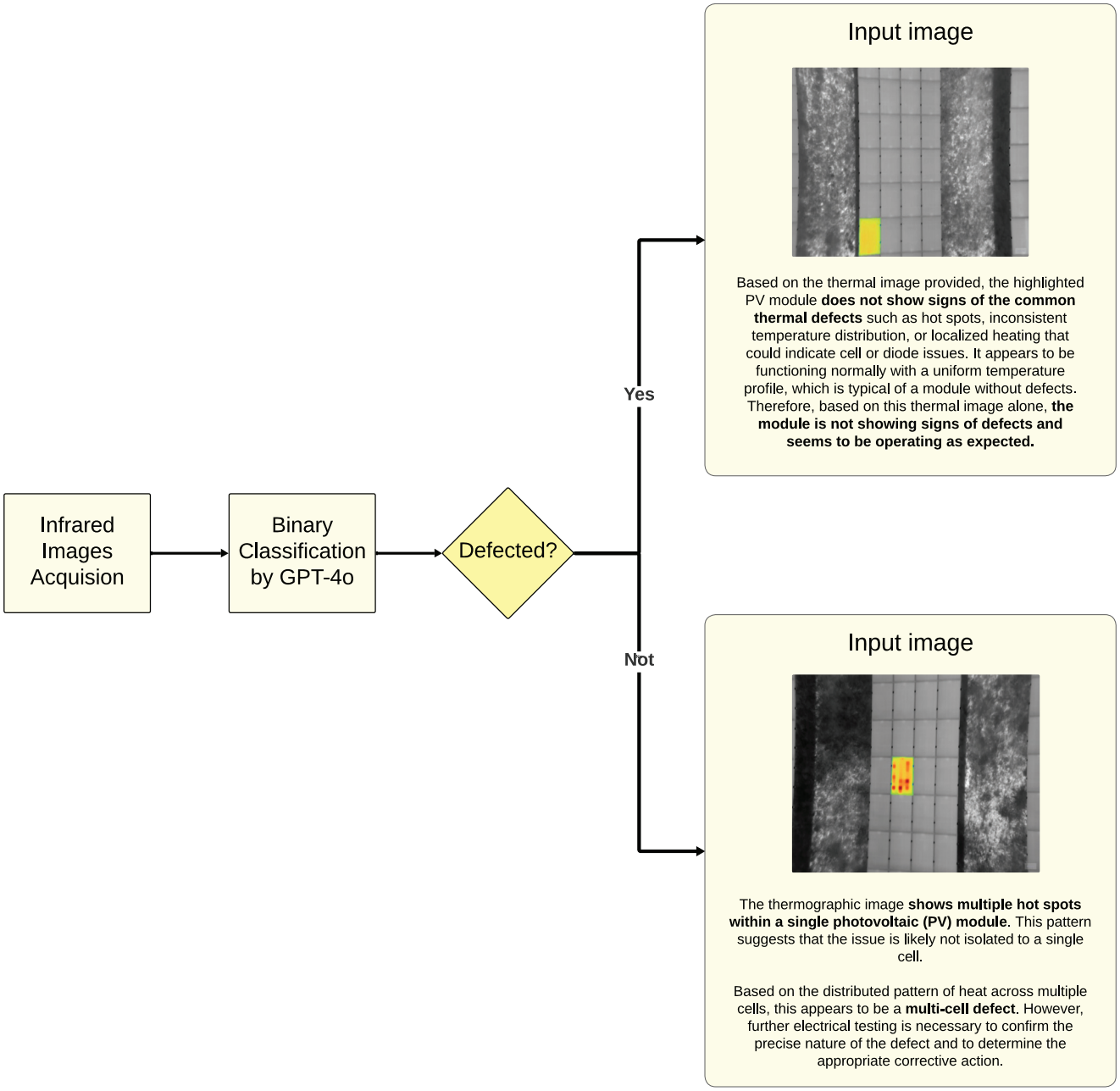

Fig. 7. GPT-4o prompt engineering for failure classification.

traditional computer vision models, it eliminates the need for extensive image preprocessing, as GPT-4o automatically handles the padding and resizing of PV module images to match its input layer specifications.

4.3.1 Different dataset, images with higher resolution

To address the poor performance of GPT-4o on low-resolution images in our initial dataset, we conducted a second experiment using high-resolution thermal images captured by unmanned drones. Our goal was to evaluate GPT-4o's performance on higher quality images while streamlining the detection and classification pipeline by eliminating preprocessing steps like module isolation, individual module detection, and perspective correction. Unlike traditional vision-based models such as ResNet, which require extensive preprocessing, we tested GPT-4o's ability to classify failures directly from raw aerial images of PV strings. The dataset used is open-source and available on Kaggle (Photovoltaic system thermography) [43]. It primarily consists of four main fault types: defective cell, multi-cell, diode, and junction box. Figure 7 shows the methodology applied for failure classification in GPT-4o with an example of the high-resolution dataset. The improved performance of the model shown in Table 5 compared to Table 4 and Table 3 highlights that higher-

Table 5. Results for a dataset of thermal images taken by a drone.

Accuracy – Classification for 4 failure classes in GPT-4o		
	Zero-shot prompting	Instructional prompting
Grey-scale	0.45	0.75
Colored	0.40	0.95

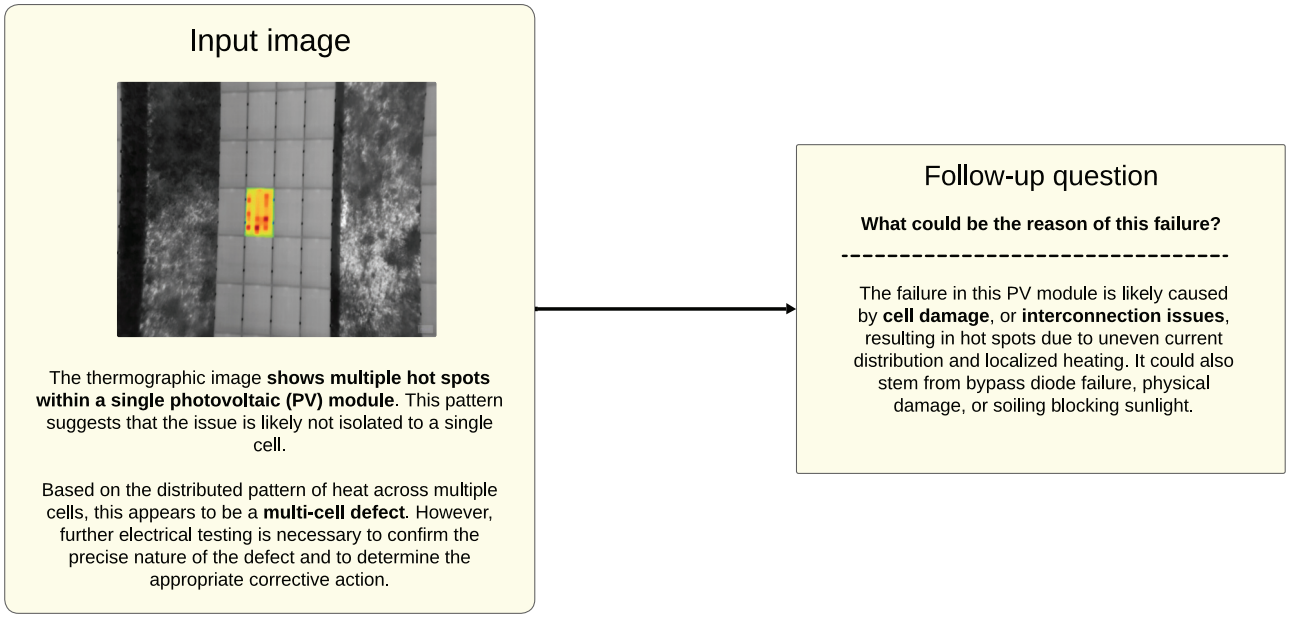

Fig. 8. Reasoning about GPT-4o classification responses.

quality thermal images can significantly enhance the accuracy of fault detection and classification using GPT-4o. The model's performance improved additionally when we switched from zero-shot to instructional prompting. This highlights the importance of effective prompt engineering for models like GPT-4o – we observed a 30% improvement in classification performance on gray-scale images and a 40% improvement on color images.

4.3.2 Visual question answering

A key advantage of using MLLM in failure classification tasks is their interactive nature. Beyond simply assigning failure labels, LLMs enable dynamic dialogue through follow-up questions, allowing users to request detailed explanations or additional context about the classification results. This capability enhances reasoning quality and provides better support for decision-making processes. Figure 8 demonstrates how GPT-4o can be used to analyze potential causes of multi-hotspot failures in semiconductor images through an interactive question-answering approach.

5 Conclusion

In this comparative study, GPT-4o and ResNet were evaluated for their effectiveness in detecting and classifying defects on PV modules using aerial infrared imagery. Each model has distinct advantages and limitations, making them suitable for different contexts.

ResNet, a convolutional neural network designed for image recognition tasks, is relatively lightweight, and has therefore demonstrated clear strengths in its low computational requirements and ease of implementation, making it a practical choice for real-time image classification tasks, particularly when working with large datasets where efficiency is paramount. However, its vision-only capabilities limit its usefulness to simple image classification tasks, restricting its ability to provide more in-depth analysis or handle more complex scenarios.

The GPT-4o, on the other hand, redefines the boundaries of fault detection in photovoltaic systems by offering a more nuanced and adaptable approach. Its ability to generate detailed image descriptions and process different types of input, including multimodal data, makes it a superior tool for advanced fault classification and analysis. This adaptability is essential for detecting and explaining subtle anomalies in infrared aerial images. However, GPT-4o is a large language model developed by OpenAI and demands substantially more computational resources due to its complex architecture resulting in slower inference times and higher operational costs [44] which makes it less suitable for real-time applications and resource-constrained environments.

In conclusion, the choice between ResNet and GPT-4o depends on the specific needs of the task in hand. For fast, efficient image classification in resource-constrained environments, ResNet is ideal. However, for more in-depth and detailed fault analysis, where interpretability and adaptability are essential, GPT-4o provides a more complete solution, pushing the boundaries of computer vision in the context of photovoltaic system fault detection.

Acknowledgments

This publication is co-funded by the European Union from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101146883, project "Supernova". Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Funding

Further, the authors acknowledge the financial support from the project PE00000021 "Network 4 Energy Sustainable Transition – NEST PNRR MUR".

Conflicts of interest

The authors declare no conflict of interest.

Data availability statement

The Python code presented in this work is shared in a GitHub repository archived in: https://github.com/evelynturri/pv-ai-detector.

Author contribution statement

Conceptualization: Sandra Gallmetzer, Mousa Sondoqah and Evelyn Turri. Methodology: Sandra Gallmetzer, Mousa Sondoqah and Evelyn Turri. Software: Sandra Gallmetzer, Mousa Sondoqah and Evelyn Turri. Writing – Original Draft Preparation: Sandra Gallmetzer, Mousa Sondoqah and Evelyn Turri. Writing – Review & Editing: Lukas Koester and Atse Louwen. Supervision: Atse Louwen and David Moser.

References

1. M. Fischer, M. Woodhouse, P. Baliozian, International technology roadmap for photovoltaics IRTPV report. Technical Specification Fifteenth Edition, May 2024, VDMA, Frankfurt, GE (2024). https://www.vdma.org/international-technology-roadmap-photovoltaic
2. U. Jahn, B. Herteleer, C. Tjengdrawira, I. Tsanakas, M. Richter, G. Dickeson, A. Astigarraga, T. Tanahashi, F. Valencia, M. Green, A. Anderson, B. Stridh, A. Lagunas, Y. Sangpongsanont, PVPS Task 13, Subtask 3: Guidelines for Operation and Maintenance of Photovoltaic Power Plants in Different Climates. Report, International Energy Agency (2022)
3. A. Stiedl, C. Sas, C. Braun et al., Operation & Maintenance – Best practice guidelines. Best practice guidelines Version 6.0, Solar Power Europe (2025)
4. W. Herrmann, G. Eder, B. Farnung, G. Friesen, M. Köntges, B. Kubicek, O. Kunz, H. Liu, D. Parlevliet, I. Tsanakas, J. Vedde, PVPS task 13: Performance, operation and reliability of photovoltaic systems – qualification of photovoltaic (PV) power plants using mobile test equipment. Report, International Energy Agency (2021)
5. J. Kuo, S.-H. Chen, C.-Y. Huang, Automatic detection, classification and localization of defects in large photovoltaic plants using unmanned aerial vehicles (UAV) based infrared (IR) and RGB imaging, Energy Convers. Manag. **276**, 116495 (2023). https://doi.org/10.1016/j.enconman.2022.116495
6. A. Fernández, R. Usamentiaga, P. de Arquer, M. Ángel Fernández, D. Fernández, J. Luis Carús, M. Fernández, Robust detection, classification and localization of defects in large photovoltaic plants based on unmanned aerial vehicles and infrared thermography, Appl. Sci. **10**, 5948 (2020). https://doi.org/10.3390/app10175948
7. S. Lindig, S. Gallmetzer, M. Herz, A. Louwen, E. Koumpli, P.S. Enriquez Paez, D. Moser, Towards the development of an optimized Decision Support System for the PV industry: A comprehensive statistical and economical assessment of over 35,000 O&M tickets, Prog. Photovolt.: Res. Appl. **31**, 1215 (2022). https://doi.org/10.1002/pip.3637
8. S. Gallmetzer, S. Lindig, M. Herz, D. Moser, Automated fixing cost estimation of photovoltaic system failures for the creation of a decision support system, Sol. RRL **7**, 2300562 (2023). https://doi.org/10.1002/solr.202300562
9. A. Ara, A. Lee, A. Sacco, A. Rahmati, A. Finch, Solarpower europe: O&m best practices guidelines version 5.0. Report, Solar Power Europe (2021). https://doi.org/10.1002/solr.202300562
10. N.R.E. Laboratory, Sandia National Laboratory, SunSpec Alliance, The S. National Laboratory Multiyear Partnership (SuNLaMP) PV Operation, and Maintenance Best Practices Working Group. Practices for Operation and Maintenance of Photovoltaic and Energy Storage Systems 3rd edition. Technical report, NREL (2018)
11. G. Schirripa Spagnolo, P. Del Vecchio, G. Makary, D. Papalillo, A. Martocchia, A review of IR thermography applied to PV systems, in *2012 11th International Conference on Environment and Electrical Engineering* (Venice, Italy, 2012). https://doi.org/10.1109/EEEIC.2012.6221500
12. IEC62446-3, Photovoltaic (PV) systems – requirements for testing, documentation and maintenance – part 3: Photovoltaic modules and plants – outdoor infrared thermography. Technical Specification IEC TS 62446-3:2017, International Electrotechnical Commission, Geneva, CH (2017). https://webstore.iec.ch/en/publication/28628
13. G. Álvarez-Tey, C. García-López, Strategy Based on Two Stages for IR Thermographic Inspections of Photovoltaic Plants, Appl. Sci. **12**, 6331 (2022). https://doi.org/10.3390/app12136331
14. U. Jahn, M. Herz, M. Köntges, D. Parlevliet, M. Paggi, I. Tsanakas, J.S. Stein, K.A. Berger, S. Ranta, R.H. French, M. Richter, T. Tanahashi, Task 13: Review on infrared and electroluminescence imaging for PV field applications (2018)
15. M. Aghaei, A. Gandelli, F. Grimaccia, S. Leva, R.E. Zich, Ir real-time analyses for pv system monitoring by digital image processing techniques, in *2015 International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)* (2015), pp. 1–6. https://doi.org/10.1109/EBCCSP.2015.7300708

16. M. Umair Ali, H. Farhaj Khan, M. Masud, K.D. Kallu, A. Zafar, A machine learning framework to identify the hotspot in photovoltaic module using infrared thermography, Sol. Energy **208**, 643 (2020). https://doi.org/10.1016/j.solener.2020.08.027
17. V. Sinap, A. Kumtepe, Cnn-based automatic detection of photovoltaic solar module anomalies in infrared images: a comparative study, Neural Comput. Appl. **36**, 1771 (2024). https://doi.org/10.1007/s00521-024-10322-y
18. Z. Nichols, Test success for robot inspector, pv magazine, November (2024). https://www.pv-magazine.com/2024/01/11/test-success-for-robot-inspector/
19. Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, L. Wang, The dawn of LMMS: Preliminary explorations with gpt-4v, arXiv:2309.17421 (2023). https://doi.org/10.48550/arXiv.2309.17421
20. J. Wu, W. Gan, Z. Chen, S. Wan, P.S. Yu, Multimodal large language models: A survey, arXiv:2311.13165 (2023). https://doi.org/10.48550/arXiv.2311.13165
21. F. Hong, J. Song, H. Meng, R. Wang, F. Fang, G. Zhang, A novel framework on intelligent detection for module defects of PV plant combining the visible and infrared images, Sol. Energy **236**, 406 (2022). https://doi.org/10.1016/j.solener.2022.03.018.
22. P. Karthikayan, V. Varshan, H. Kattamuri, U. Jayaraman, Explainable AI: Comparative analysis of normal and dilated resnet models for fundus disease classification, arXiv:2407.05440 (2024). https://doi.org/10.48550/arXiv.2407.05440
23. M. Köntges et al., Review of failures of photovoltaic modules, Report IEA-PVPS (2014)
24. R.H.F. Alves, G.A. de Deus Júnior, E.G. Marra, R. Pinto Lemos, Automatic fault classification in photovoltaic modules using convolutional neural networks, Renew. Energy **179**, 502 (2021). https://doi.org/10.1016/j.renene.2021.07.070
25. L. Koester, S. Lindig, A. Louwen, A. Astigarraga, G. Manzolini, D. Moser, Review of photovoltaic module degradation, field inspection techniques and techno-economic assessment, Renew. Sustain. Energy Rev. **165**, 112616 (2022). https://doi.org/10.1016/j.rser.2022.112616
26. Á.H. Herraiz, A. Pliego Marugán, F.P. García Márquez, Photovoltaic plant condition monitoring using thermal images analysis by convolutional neural network-based structure, Renew. Energy **153**, 334 (2020). https://doi.org/10.1016/j.renene.2020.01.148
27. C. Dunderdale, W. Brettenny, C. Clohessy, E. Van Dyk, Photovoltaic defect classification through thermal infrared imaging using a machine learning approach, Prog. Photovolt.: Res. Appl. **28**, 12 (2019). https://doi.org/10.1002/pip.3191
28. G. Cipriani, A. D'Amico, S. Guarino, D. Manno, M. Traverso, V. Di Dio, Convolutional neural network for dust and hotspot classification in PV modules, Energies **13**, 6357 (2020). https://doi.org/10.3390/en13236357
29. D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, D. Yu, Mm-llms: Recent advances in multimodal large language models, arXiv:2401.13601 (2024). https://doi.org/10.48550/arXiv.2401.13601
30. R. Desislavov, F, Martínez-Plumed, J. Hernández-Orallo, Trends in Ai inference energy consumption: beyond the performance-vs-parameter laws of deep learning, Sustain. Comput.: Inform. Syst. **38**, 100857 (2023). https://doi.org/10.1016/j.suscom.2023.100857
31. J.-O. Schneppat, Pre-trained Models (2019). https://schneppat.com/pre-trained-models.html
32. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
33. O. AI, J. Achiam et al., Gpt-4 technical report (2024). https://arxiv.org/abs/2303.08774
34. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
35. B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering in large language models: a comprehensive review, arXiv:2310.14735 (2024). https://doi.org/10.48550/arXiv.2310.14735
36. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q.V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, arXiv:2201.11903 (2022). https://doi.org/10.48550/arXiv.2201.11903
37. J. Wu et al., Visual prompting in multimodal large language models: a survey, arXiv:2409.15310 (2024). https://doi.org/10.48550/arXiv.2409.15310
38. S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, arXiv:2306.13549 (2024). https://doi.org/10.48550/arXiv.2306.13549
39. X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, P. Lucas, H. Sauzéon, P.-Y. Oudeyer, Selecting better samples from pre-trained llms: case study on question generation, arXiv:2209.11000 (2022). https://doi.org/10.48550/arXiv.2209.11000
40. Y. Ni, S. Jiang, X. Wu, H. Shen, Y. Zhou, Evaluating the robustness to instructions of large-language-models, arXiv:2308.14306 (2023). https://doi.org/10.48550/arXiv.2308.14306
41. M. Millendorf, E. Obropta, N. Vadhavkar, Infrared solar module dataset for anomaly detection, in *ICLR 2020* (2020)
42. D. Jordan, T. Silverman, J. Wohlgemuth, S. Kurtz, K. VanSant, Photovoltaic failure and degradation modes: PV failure and degradation modes, Prog. Photovolt.: Res. Appl. **25**, 04 (2017). https://doi.org/10.1002/pip.2866
43. G. Marcos, System thermography (2023). https://www.kaggle.com/code/marcosgabriel/dataset-intro-photovoltaic-system-thermography
44. GPT-4: Quality, Performance & Price Analysis. https://artificialanalysis.ai/models/gpt-4?utm_source=chatgpt.com

Cite this article as: Sandra Gallmetzer, Mousa Sondoqah, Evelyn Turri, Lukas Koester, Atse Louwen, David Moser, Redefining failure detection in PV Systems: a comparative study of GPT-4o and ResNet's computer vision in aerial infrared imagery analysis, EPJ Photovoltaics **16**, 23 (2025), https://doi.org/10.1051/epjpv/2025010