DEPARTMENT OF CHEMICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS
CHENNAI-600 036

DEEP LEARNING APPROACHES IN SUSTAINABLE ENERGY

Predictive Maintenance of Solar PV Panel by Fault Detection & Percentage of Damaged Identification

A THESIS REPORT

Submitted by

Aloy Banerjee

(Roll No- CH22M503)

For the award of the degree

of

MASTER OF TECHNOLOGY

IN INDUSTRIAL AI

Under the guidance of **Dr. Apurba Das**

THESIS CERTIFICATE

This is to certify that the thesis titled **DEEP LEARNING APPROACHES IN SUSTAINABLE ENERGY: Predictive Maintenance of Solar PV Panel by Fault Detection & percentage of damaged identification**, submitted by **Aloy Banerjee (CH22M503),** to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology,** is a bonafide record of the research work, done by him under my supervision. The contents of this thesis, in whole or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

DR. APURBA DAS
Research Guide
Cognitive AI Head
TATA Consultancy Services

Place: Chennai
Date: 31st July 2024

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to my mentor, **Dr Apurba Das,** for his invaluable guidance, encouragement, and expert advice throughout my M.Tech project. His insights and expertise were helpful in the completion of this research, and his support has been a constant source of motivation.

I am also immensely grateful to my employer**, TATA Consultancy Services,** for sponsoring this M.Tech course. The opportunity to advance my education and skills through this sponsorship has been a pivotal step in my academic and professional growth. The support and understanding of my colleagues at **TATA Consultancy Services** have also been greatly appreciated, enabling me to balance my work and study commitments effectively.

This research would not have been possible without the professors of **IIT Madras** their suggestions and valuable feedback on this work. I sincerely thank all the professors who were directly/indirectly involved with the course.

Furthermore, I extend my heartfelt appreciation to my family and friends for their unwavering support and understanding throughout this journey. Their love and encouragement provided a constant source of strength during challenging moments.

ABSTRACT

Building upon our previous work on solar power forecasting, which introduced a novel LSTM-based XGBoost ensemble network for enhanced prediction accuracy, this study extends the research by addressing the critical aspect of photovoltaic (PV) panel fault detection. Recognizing that PV panel faults, such as soiling, bird-drop, physical-damage, and electrical defects etc., can significantly impact the efficiency and output of solar power systems Our new approach can integrates fault classification and segmentation into the existing framework for early fault detection and categorization, enhancing overcall efficacy.

This research presents a novel method that uses machine learning and image analysis to detect faults in PV panels early and accurately. We combine machine learning, transfer learning, and segmentation algorithms to classify and segment faults, using a new dataset of labelled images showing different fault types. Our system not only identifies and classifies faults but also measures the affected areas, offering a thorough assessment of panel health.

By incorporating these enhancements, we aim to optimize maintenance routines and ensure the reliability of solar power systems. The integration of image analytics with our previous predictive model facilitates a robust and holistic approach to solar power forecasting and maintenance, addressing the inherent challenges of PV panel faults.

Our preliminary results in solar power forecasting have demonstrated a substantial 49.50% improvement over the pure LSTM method. By now integrating significant advancements in fault detection accuracy, we can develop a robust and comprehensive system for solar power plants. This enhanced system not only advances the field of solar energy prediction but also establishes a solid foundation for future research on predictive maintenance in renewable energy systems.

KEYWORDS: PV Panel Fault Detection, Deep Learning, Image Analytics, Predictive Maintenance, Renewable Energy Systems

ABBREVIATION

- **LSTM -** Long Short-Term Memory
- **ANN -** Artificial Neural Network
- **RNN -** Recurrent Neural Networks
- **CNN -** Convolutional Neural Network
- **ML** - Machine Learning
- **DL** - Deep Learning
- **PV** - Photovoltaic Panel

LIST OF SYMBOLS

- $\boldsymbol{X}$: Pixel vector in an image
- $\boldsymbol{D} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_L}\}$: Set of fault categories
- $\boldsymbol{f}(\boldsymbol{x}): \boldsymbol{X} \to \boldsymbol{D}$: Classification function mapping pixel vector X to fault category in D
- $\boldsymbol{S}$: Segmented area of faults
- **A:** Total area of the solar panel
- **Percentage of Fault Area** $= \left(\frac{S}{A}\right) \times \mathbf{100}\%$: Percentage of the fault area
- $\boldsymbol{S} = \frac{1}{2}\left|\sum_{i-1}^{n-1}(x_i y_{i+1} - y_i x_{i+1}) + (x_n y_1 - y_n x_1)\right|$ **:** Shoelace formula for calculating the area of a fault
- $\boldsymbol{A}$: Area of the fault
- $\boldsymbol{A}_{\mathbf{max}}$: Maximum area of the solar panel
- $\boldsymbol{F1}, \boldsymbol{F2}, \dots, \boldsymbol{Fn}$: Classified fault
- $\boldsymbol{A1}, \boldsymbol{A2}, \dots, \boldsymbol{An}$: Calculated fault areas
- $\boldsymbol{X} = [\boldsymbol{P}, \boldsymbol{C}, \boldsymbol{V}]$; Features including pixel values P, color channel intensity C , and vertices V
- $\boldsymbol{f}$: Classification model with parameters θ
- $\boldsymbol{g}$: Area estimation model with parameters θ
- $\hat{\boldsymbol{F}} - \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\theta})$: Predicted fault classification
- $\hat{\boldsymbol{A}} - \boldsymbol{g}(\boldsymbol{X}; \boldsymbol{\theta})$: Predicted fault area
- $\mathbf{min}\sum_{t-1}^{T}(\boldsymbol{L}(\boldsymbol{F}, \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\theta})) + \boldsymbol{\alpha} \times \boldsymbol{E}(\boldsymbol{A}, \boldsymbol{g}(\boldsymbol{X}; \boldsymbol{\theta})) - \boldsymbol{\lambda} \times (\mathbf{1} - \boldsymbol{R}))$: Optimization goal
- **Li**: Loss function for classification

- **E**: Error function for area estimation
- **R**: Recall
- $\boldsymbol{\alpha}$: Weight for area estimation error
- $\boldsymbol{\lambda}$: Balance parameter between loss minimization and recall maximization
- $\boldsymbol{\theta}^* = \mathbf{arg\ min}\boldsymbol{\theta}\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i-1}}^{\boldsymbol{N}}\left(\boldsymbol{L}\left(\boldsymbol{y_i}, \boldsymbol{f}(\boldsymbol{X_i}; \boldsymbol{\theta})\right) + \boldsymbol{\alpha} \times \boldsymbol{E}\left(\boldsymbol{a_i}, \boldsymbol{g}(\boldsymbol{X_i}; \boldsymbol{\theta})\right) - \boldsymbol{\lambda} \times (\mathbf{1} - \boldsymbol{R})\right)$: Optimal parameter for minimizing loss and maximizing recall
- **N**: Number of training samples
- $\boldsymbol{y_i}$: Actual classification label
- $\boldsymbol{a_i}$: Actual fault area

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

In the quest for sustainable energy solutions, solar power stands out due to its abundant availability and environmental benefits. The efficient harnessing and transformation of solar energy into electrical power heavily depends on the optimal performance of photovoltaic (PV) panels. However, the effectiveness of these panels can be compromised by various faults such as microcracks, soiling, physical damage, and electrical defects. Building on our previous research, which introduced a novel LSTM-based XGBoost ensemble network for improved solar power forecasting, this study extends to incorporate advanced image analytics for fault detection in PV panels. By integrating these techniques, we can enhance the accuracy of solar power predictions and the reliability of solar power systems.

1.2 Problem Statement

The global expansion of photovoltaic (PV) power installations demands the implementation of sophisticated predictive maintenance techniques to maximize operational efficiency and extend system longevity. Traditional monitoring approaches, primarily reactive in nature, frequently overlook incipient and underlying faults within solar panels, precipitating considerable declines in energy output and escalating maintenance expenditures. Such systems are often compromised by the inherent variability in environmental factor and physical obstructions causing shading. These elements significantly challenge the consistency of fault detection processes. Recognizing and addressing these faults is crucial not only for maintaining system health but also for enhancing the accuracy of integrated solar power generation forecasting models. This dual approach ensures both the sustained functionality of the PV systems and the reliability of energy production forecasts.

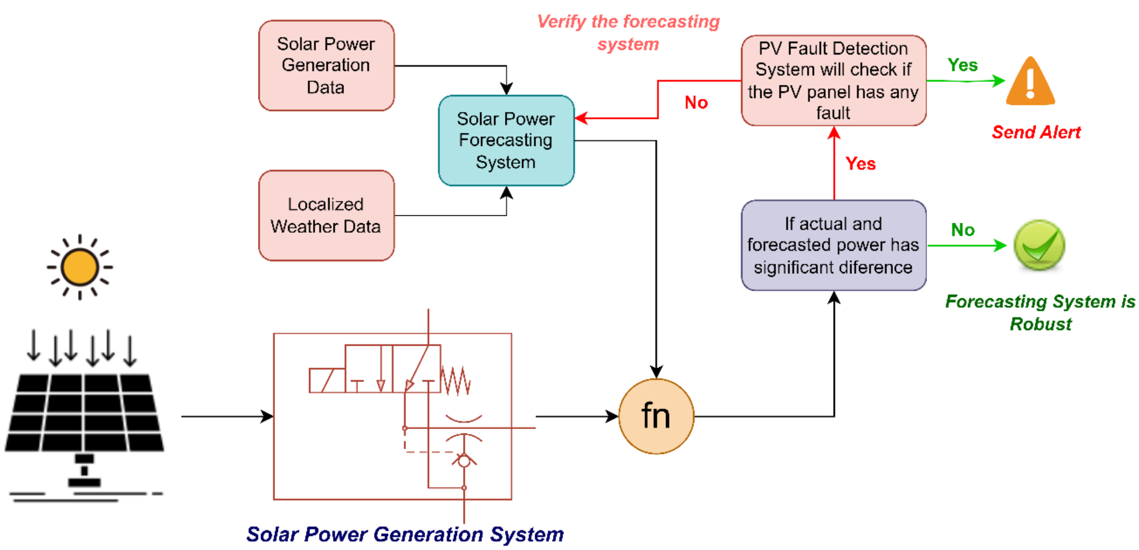
1.3 Background study

The initial phase of our research concentrated on the development of predictive models for solar power generation through the utilization of advanced deep learning techniques. Specifically, we employed an LSTM-based XGBoost ensemble network, which demonstrated a substantial improvement in prediction accuracy compared to standalone models. However, despite these advancements, the issue of photovoltaic (PV) panel faults remained unaddressed, a critical factor influencing the efficiency of solar energy systems. To address this problem, we have done thorough review of existing work and here what we have found is explained in a very high level in following paragraph and in detailed further in Literature Review section.

Recent studies have underscored the potential of convolutional neural networks (CNNs) in various image analysis tasks, including medical imaging and satellite data interpretation. In the domain of PV fault detection, CNNs have yielded promising results. For instance, Shihavuddin et al. (2017)[22] illustrated that CNNs could accurately identify various types of faults, such as cracks and hotspots, from thermal images of PV panels. Pierdicca et al. (2018)[23] conducted a comprehensive review of image pattern recognition applied to PV systems, highlighting the effectiveness of CNNs in fault detection and classification tasks. Additionally, Amaral et al. (2019)[24] proposed a method for fault diagnosis in PV tracking systems utilizing image processing techniques and principal component analysis (PCA), achieving significant improvements in detection accuracy. Furthermore, Zhang et al. (2022)[25] proposed an ensemble approach that combines CNNs with decision trees, achieving superior fault classification performance relative to single-model strategies. These advancements indicate that similar methodologies could be effectively applied to detect and classify faults in PV panels, thereby enhancing the overall reliability and efficiency of solar energy systems.

1.4 Motivation

The primary motivation behind this research is the need for a comprehensive approach to solar power forecasting that considers both power generation and system health. The efficiency of solar power systems is highly dependent on the condition of PV panels. Faults such as microcracks, soiling, and electrical issues can significantly degrade performance. By integrating advanced image analytics for fault detection into our predictive models, we aim to provide a more accurate and reliable forecasting system. This integration not only enhances the predictive accuracy but also supports proactive maintenance strategies, thereby extending the lifespan of PV panels and ensuring consistent energy production. To be system should look like below fig1.1 which increase the efficacy of the overcall power supply system.



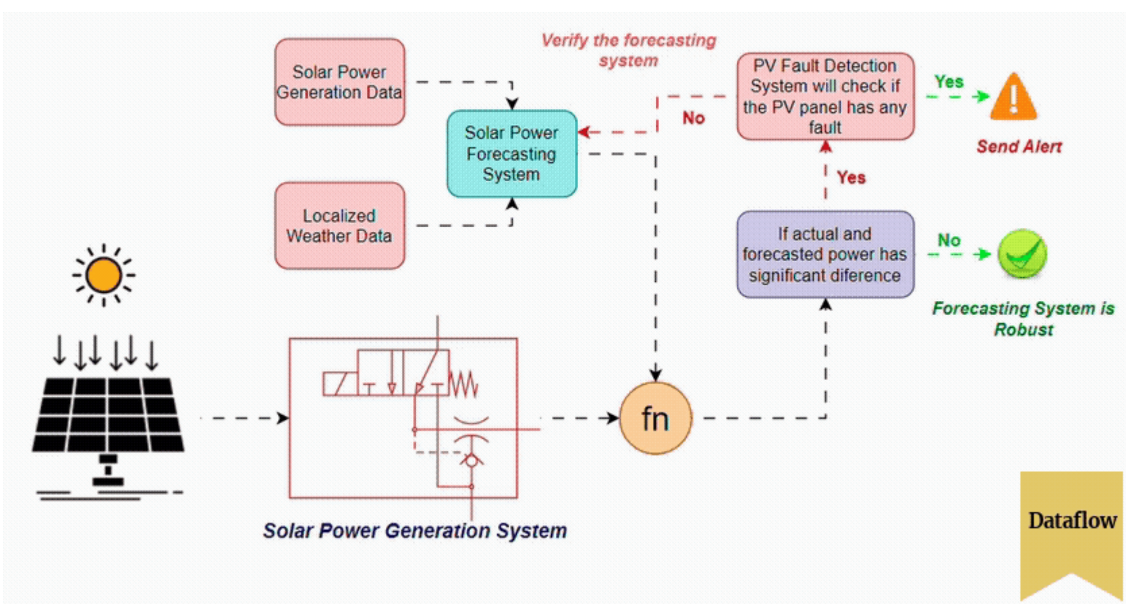


Figure 1.1: PV Panel Power Supply Monitoring and Fault Detection System

1.5 Research Objectives

The aim of this research is to harness machine or deep learning technologies to improve predictive maintenance strategies for solar PV panels. Specifically, the project will:

- **Develop a Model:** Construct and validate a predictive maintenance model that leverages both and tabular fault information to detect and classify panel faults & electrical faults more accurately.
- **Enhance Fault Detection Accuracy:** Implement a novel architecture to enhance the system's ability to identify various fault types under diverse operational conditions.
- **Fault Area Measurement**: Perform segmentation of the PV panel fault and try to measure the area of the segmented faults using Shoelace formulation. This mathematical approach helps in quantifying the size of the faults detected in the PV panels, providing an additional metric for assessing the severity of the faults.
- **Evaluate Model Performance**: Assess and compare the efficacy of the hybrid model against traditional monitoring methods and models using single data sources in terms of various metrices.

1.6 Approached to be used in Study

Our primary goal is to find out the faults and classify it and later try to segment it for details understanding. To achieve this target this thesis subsequently divides the task into two major parts. In first part we are mainly focused on the nuance of fault classification and followed by a segmentation of faults in photovoltaic panel,

1.6.1 Fault Classification

To achieve the desired outcomes, this study employs a new approach to predictive maintenance of photovoltaic (PV) panels. It integrates advanced image analytics and machine learning techniques, after evaluating several convolutional neural networks (CNNs) and noting their significant training times. The proposed methodology focused on a hierarchical ensemble classification system which is enhanced by dynamic feature extraction and dimensionality reduction. Additionally, incorporate tabular fault data and introduce feedback mechanism alongside incremental learning through partial retraining on new incoming tabular fault data, while incorporating only feedback mechanism for image data. This ensures that the model continuously improves with the latest information & user feedback.

The proposed architecture begins with converting images to grayscale, using a pretrained model for feature extraction. Principal Component Analysis (PCA) is then used to retain significant variance in the image data. The dataset is split into training, testing, and validation sets, with incremental PCA employed to assess local data characteristics dynamically. The core of the classification process involves training a Support Vector Machine (SVM) classifier on the PCA-transformed data. Misclassified instances are further processed using a Naive Bayes classifier, with data scaled according to feature weights from the initial PCA. This hierarchical approach allows for corrections and improvements in overall predictions.

Furthermore, the model leverages tabular fault data for fault classification and employs a decision level fusion method to develop a multimodal approach. Additionally, we incorporate feedback mechanism and incremental learning techniques for tabular fault data by partially fitting on new data and adapting based on feedback from image data. This strategy ensures the model continuously improves with new information.

By combining these methods, the thesis aims to develop a robust and comprehensive system for PV panel fault detection and maintenance, significantly improving the reliability and efficiency of solar power systems.

1.6.2 Fault Segmentation

The second part of the study will focus on the segmentation of faults in photovoltaic panels to provide a detailed understanding of the fault's nature and extent. This will be achieved by implementing a segmentation algorithm, preferably U-Net, to accurately delineate the fault regions in the PV panel images. The chosen segmentation algorithm will aim to separate faulty areas from the rest of the panel, ensuring precise segmentation.

Following successful segmentation, the Shoelace algorithm will be applied to calculate the area of the segmented faults. This mathematical approach will provide an accurate measurement of the fault area, which is crucial for assessing the severity and impact of the faults on the PV panels' performance.

This two-step process of segmentation and area calculation will enhance the overall fault detection system, offering a more comprehensive analysis of PV panel health. The integration of a suitable segmentation algorithm and the Shoelace algorithm for area calculation aims to provide detailed insights into the fault characteristics, thereby improving maintenance strategies and ensuring the efficiency and reliability of solar power systems.

1.7 Organization of the thesis

Subsequent sections of the thesis are structured as follows:

In ***Chapter 2,*** related works in the relevant fields.

In ***Chapter 3,*** explain the problem statement and formulation.

In ***Chapter 4,*** provide an explanation of the dataset used in this study.

In ***Chapter 5,*** provide the proposed approach and the entire workflow of the system.

In ***Chapter 6,*** explain the detailed experimental results.

In ***Chapter 7*****,** conclude our observation and explain the future pathway for this research.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

The burgeoning field of predictive maintenance for solar photovoltaic (PV) panels has garnered significant attention, driven by the imperative to enhance the efficiency and longevity of solar energy systems. Traditional maintenance practices, which often rely on periodic inspections and reactive repairs, are being increasingly supplemented or replaced by advanced predictive maintenance strategies leveraging deep learning techniques. This section presents a comprehensive examination of the current body of investigation that pertains to the utilization of deep learning methods in the prediction of maintenance for photovoltaic (PV) systems. Several studies have highlighted the effectiveness of CNNs in analyzing infrared thermography images and electroluminescence images of solar panels to identify hot spots, micro-cracks, and other anomalies indicative of faults. These models have been trained on large datasets to recognize complex patterns associated with various types of defects, significantly improving detection accuracy. The literature also highlights the importance of accurately quantifying the percentage of damage in PV panels.

2.2 Fault Detection and Damaged Identification in PV system

Fault detection and damage identification in photovoltaic (PV) systems are critical for maintaining system efficiency, safety, and longevity. The literature in this field encompasses a variety of approaches, including electrical parameter monitoring, thermal imaging, and machine learning algorithms. A variety of writers have conducted research on fault detection and damage diagnosis in PV systems and have presented below.

In their study, Yousif et.al (2024) [1] introduced a new deep learning model that combines human and automatic feature extraction methods to improve the precision of PV photo classification. This research made a significant contribution by developing a hybrid model that enhances the feature vector used in deep neural networks by using the histogram of oriented gradient (HoG) of PV images. The experimental results show that our strategy outperforms six other current methods that use either the same or different underlying deep learning model.

Kaitouni et.al (2024) [2] conducted a study that specifically aimed to identify and diagnose malfunctions in decentralized solar PV systems inside metropolitan areas. This strategy integrates a hybrid model-based and data-driven methodology, including monitoring and inspection tools such as Remote Sensors (RS) and a real-time solar production monitoring system. The use of the Decision Tree (DT) concept in intricate dynamic systems has shown to be effective in improving the operational parameters of energy systems. This is accomplished by evaluating the spatial and temporal energy efficiency of dispersed solar PV systems on an urban scale, adjusting to the changing surroundings, and identifying abnormalities that suggest deterioration. The main aim of this study is to improve comprehension of the research on monitoring the state of solar systems in urban environments, specifically by addressing anomalies such as the steady degradation of PV modules over time. The findings of this research indicate that the state and monitoring of photovoltaic (PV) systems may be precisely detected and successfully tracked by integrating real-time monitoring with remote aerial sensing. In their study, Duranay et.al (2023) [3] shown that Photovoltaic panels play a crucial role in the energy conversion process, making a substantial contribution to the sustainability and environmental compatibility of solar energy. However, any flaws in solar panels may greatly hinder energy generation, making it crucial to promptly identify them. The feasibility of using infrared photographs of solar modules for flaw detection in photovoltaic panels via the use of deep learning algorithms. This breakthrough would enhance the sustainability and effectiveness of solar energy systems. The dataset employed consisted of twenty thousand images collected using infrared solar modules. The gathered photos were categorized into twelve distinct groups: vegetation, cell, cell-multi, fracture, no anomaly, offline-module, shadowing, soiling, diode, hot spot, hot spot-multi, and vegetation. The approach was developed using the Efficientb0 concept. The NCA feature selector was used to retrieve a significant number of 17,000 features from the example model. Afterwards, a Support Vector Machine (SVM) classifier was used to do the classification.

Mustafa et.al (2023) [4] emphasized the need of precise fault classification and location in order to achieve maximum performance of a photovoltaic (PV) system. PV systems are vulnerable to several types of fault conditions, such as string-to-string (SS), string-to-ground (SG), and open-circuit faults (OC), because to the ambient factors in which they function. If left unaddressed, these imperfections might lead to the depletion of power and, eventually, problems in the system. This study presents a technique for mitigating the effects of these consequences by utilizing advanced deep learning (DL) methods such as convolutional neural networks (CNN), long short-term memory (LSTM), and bi-directional long short-term memory (Bi-LSTM) networks to identify, categorize, and locate SS, SG, and OC issues. The proposed strategy reduces the necessary number of sensors per string by 50% compared to previous methods described in the literature. Furthermore, it facilitates improved scalability and the generation of a dataset with several labels using a decoupled modeling method. The fault classification has an accuracy rate of 99.94%, while the fault location has an accuracy rate of 99.54%. In addition, a comparison is made between the proposed deep learning (DL) algorithms and typical machine learning algorithms using many evaluation metrics. The efficacy of the proposed methodology is proved by the analysis of diverse datasets and the deliberate introduction of noise. The results demonstrate a strong degree of reliability and highlight the effectiveness and scalability of the proposed multi-output deep learning approach. Constantin et.al (2023) [5] conducted a study which is focusing on the usage of solar energy systems as a remedy for the rising global energy needs, which contribute to an excess of greenhouse gases and the subsequent mitigation of climate crises and global warming. Accurate execution of maintenance and operation tasks is essential for accurately determining the return on investment and guaranteeing the long-lasting and reliable performance of solar systems. Currently, the existing standards mostly focus on fundamental operations, such as eye examination and planned maintenance schedules. Moreover, thermography has the ability to distinguish between different types of faults that can occur in solar panels, such as bypass diode activation, junction box temperature increase, single hot spot, and multiple hot spots. It can also identify the specific characteristics and consequences associated with each of these fault types. Regular scheduled maintenance It is advisable to use an AI-powered thermal imaging software to detect faulty solar panels that need to be replaced or repaired, thereby enhancing the system's efficiency.

Eldeghady et.al (2023) [6] proposed the integration of heuristic particle swarm optimization and Back Propagation Neural Network (BPNN-PSO) techniques enhances convergence and prediction precision in the context of defect diagnostics for photovoltaic (PV) array systems. The technique integrates the classification and prediction capabilities of deep learning with the optimization capabilities of particle swarms in order to locate the optimal solution. From the output of the PV array, several parameters are extracted in order to identify and diagnose defects. An analysis is conducted to compare the outcomes produced by the back propagation neural network technology when utilized in isolation and with the addition of the back propagation heuristic combination technique. The BP-PSO technique demonstrates convergence in the training phase after a mere 250 steps, while the back propagation method requires 350 steps to converge. The BP algorithms produce an approximation of 87.8% accuracy in predictions, whereas the suggested BP-PSO techniques produced 95% accurate predictions. Both the convergence of the simulation and the precision of defect diagnosis prediction in the PV system were demonstrated to be significantly enhanced by employing the back propagation heuristic combination technique.

Hassan et.al (2023) [7] examined by means the identification of flaws in solar photovoltaic (PV) modules is critical for ensuring their optimal performance and long-term dependability. The advancement of convolutional neural networks (CNNs) has substantially enhanced the capability of fracture detection, providing greater precision and efficacy in comparison to conventional approaches. A comprehensive analysis and comparison of CNN-based solar PV module fracture detection techniques. An assortment of convolutional neural network (CNN) architectures are available, encompassing custom-built networks, ensemble learning approaches, pre-trained models, and data augmentation techniques. The discourse additionally addresses obstacles pertaining to the scale of the datasets, the applicability of findings to different solar panels, the interpretability of CNN models, and real-time detection. The assessment also suggests paths that might entail quicker computation rates, better interpretability of the model, and more and more varied datasets.

Van et.al (2023) [8] ascertain fault in photovoltaic (PV) systems have a significant effect on the efficiency and reliability of solar power, resulting in energy losses and increased costs. Current automated fault diagnosis technologies are only economically feasible for large-scale systems due to the exorbitant cost of sensors. To address these constraints, we propose a defect diagnosis model that leverages graph neural networks (GNNs). This model is specifically developed to monitor a group of photovoltaic (PV) systems by evaluating and comparing the data on their current and voltage generation over the last 24 hours. This method allows for the monitoring of photovoltaic (PV) systems without the use of sensors, since it obtains hourly measurements of the produced current and voltage via the inverters of the PV systems. Extensive testing was conducted by simulating six different photovoltaic (PV) systems in Colorado, using six years' worth of genuine weather data. The GNN has high efficacy in detecting and categorizing first occurrences of 6 prevalent mistakes, even in the presence of substantial variations in module count, module type, orientation, location, and other relevant parameters. The GNN attains an accuracy of 84.5% in the absence of weather data, and 87.5% while including satellite weather estimates. It surpasses two cutting-edge PV fault diagnosis techniques. Moreover, the results suggest that GNN has the capacity to transfer its expertise to photovoltaic (PV) systems that it was not explicitly taught on. Furthermore, it exhibits exceptional precision even in the presence of multiple simultaneous failures in several PV systems.

Pathak et.al (2022) [9] said that the output of solar energy has seen exponential growth. Sophisticated and robust condition monitoring systems are important to ensure the reliability of photovoltaic solar power facilities as they expand and become more complex. A novel method is proposed for detecting defects in solar panels, using the analysis of thermal pictures captured by a thermographic camera. Using two sophisticated convolutional neural network models, the first model is responsible for identifying the kind of defect affecting the panel, while the second model is responsible for determining the specific area of the damaged panel. When assessing different classification models, the recommended approach use the F1 score as a metric. Among these models, the ResNet-50 transfer learning model has the highest score of 85.37%. Mean Average Precision is used to evaluate object detection models, and Faster R-CNN earns the highest score of 67%. A method for improving the effectiveness of early identification and localization of defects in solar panels by reducing the need for human work.

Haidari et.al (2022) [10] establish that there is a fast worldwide spread of efforts to decrease air pollution by using renewable energy sources, including solar energy. Photovoltaic power facilities consist of a significant number of photovoltaic modules, which are renewable sources of solar energy and need regular monitoring. However, traditional methods of analyzing these modules are not feasible and might be dangerous. An algorithmic technique based on deep learning for the study of solar power systems. Two specific types of defects were examined in solar powerhouses: hotspot and hot substring. These flaws occur more often in the solar generator. The datasets used in this study consist of thermal pictures of solar modules captured from both aerial and ground perspectives. Various statistical criteria were used to assess the efficacy of the constructed network, including F1 score, accuracy, and precision. The complete network achieved a classification accuracy of 0.98, and the collected data underwent review.

Table 2.1 depict the comparison of reviewed literature on fault detection and damaged identification in PV system. This table shows various authors findings and their future work have been examined as shown below.

Table 2.1 Comparison of reviewed literature

Authors & Year	**Methods**	**Key Findings**	**Future Work**
Yousif et. al (2024) [1]	Hybrid model	The proposed model outperforms as compared to other six models.	In future, it will increase their reliability and ensure their optimal performance
Kaitouni et. Al (2024) [2]	Decision Tree	Monitoring of PV system may be precisely detected.	Furthermore, it will improve the quality of solar cell production and increase the efficiency of solar energy generation
Duranay et.al (2023) [3]	Infrared images, Efficientb0 model, NCA feature selector, SVM classifier	Identified 12 categories of defects in photovoltaic panels. Used Efficientb0 model with SVM classifier.	Optimize sustainability and efficacy of solar energy systems through defect detection.
Mustafa et. al (2023) [4]	Hybrid deep learning	The fault classification has an accuracy rate of 99.94%, while the fault location has an accuracy rate of 99.54%.	Further Improving scalability and the generation of a dataset with several label.
Constantin et.al (2023) [5]	Focus on maintenance of solar energy systems using thermography, AI-based thermal	Thermography can differentiate among various fault types in solar panels. Recommends preventative maintenance	Continued research on AI-based techniques for efficient maintenance of solar systems.

	imagery processing program	based on AI for defect detection.	
Eldeghady et.al (2023) [6]	Combination of heuristic particle swarm optimization and Back Propagation Neural Network (BPNN-PSO)	BP-PSO technique demonstrates improved convergence and 95% accuracy in defect diagnosis predictions for PV systems.	Further exploration of AI techniques for defect diagnostics in PV array systems.
Hassan et.al (2023) [7]	Comprehensive analysis and comparison of CNN-based solar PV module fracture detection techniques	CNN-based techniques provide greater precision and efficacy in fracture detection. Address challenges like dataset scale, model interpretability, and real-time detection.	Explore paths for faster computation rates, better model interpretability, and diverse datasets for CNN-based fracture detection techniques.
Van et.al (2023) [8]	GNN	The GNN attains an accuracy of 84.5% in the absence of weather data, and 87.5% while including satellite weather estimate.	Furthermore, it exhibits exceptional precision even in the presence of multiple simultaneous failures in several PV systems.
Pathak et.al (2022) [9]	Heat images processed with thermo graphic camera, ResNet-50 transfer learning model, Faster R-CNN	ResNet-50 achieves 85.37% F1 score for defect classification; Faster R-CNN attains 67% Mean Average Precision for object identification.	Enhance efficiency of early detection and defect localization in solar panels through improved classification models.
Haidari et.al (2022) [10]	Deep learning-based algorithmic approach,	Investigated hotspot and hot substring defects with	Further research and evaluation of deep

	aerial and terrestrial thermal images	a classification accuracy of 0.98.	learning-based algorithms for routine inspection of solar modules.

2.3 Solar PV system using Deep Learning Approaches

Deep learning techniques have generated considerable interest for their capacity to improve the performance, reliability, and problem detection of solar PV systems. Deep learning methods, namely convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are used to evaluate large quantities of data produced by PV systems. These models have the ability to recognize patterns and irregularities that are often imperceptible using conventional approaches, resulting in more precise forecasts of power generation, efficiency, and possible malfunctions. This section presents a thorough examination of solar PV systems using deep learning algorithms and their results.

In their study, Kumar et.al (2024) [11] created a very efficient model for predicting solar irradiance. They then included these predictions into the planned hybrid battery switching stations to optimize the use of solar electricity. Each year, new ideas and approaches are used to improve the accuracy of the model, with the main goal of minimizing the variability of predictions. This study comprehensively investigates several deep-learning methodologies to precisely predict solar irradiance within a certain time period. Analyzing time series data requires the use of real-time data. For this investigation, three methods have been utilized: the statistical ARIMA model, the LSTM-based RNN approach, and the Dual Attention-based Recurrent Neural network. The models are implemented using Jupyter Notebook and the Python programming language. Several error measures have been compared in studies to assess the efficiency and evaluate the performance of real-time data models. This study provides evidence that the LSTM model surpasses other models in predicting solar power, as indicated by its superior error metrics: a mean squared error (MSE) of 0.0091, a mean absolute error (MAE) of 0.0525, a root mean squared error (RMSE) of 0.0953, and a mean squared logarithmic error (MSLE) of 0.0047. Maciel et.al (2024) [12] introduced an innovative hybrid prediction method that is well-suited for projecting solar irradiation in the short-term. The suggested methodology employs a set of image processing measures to extract characteristics from all-sky photos. These characteristics are then used as input in machine learning-based prediction models. The set of all-sky image metrics, acquired by image processing, demonstrate complementary characteristics of the sky that lead to an average accuracy that is 30% higher compared to the use of traditional meteorological data. The testing results of the hybrid prediction approach, which integrates Artificial Neural Network and Light Gradient Boosting Machine, showed an average forecast accuracy that was 17.5% superior than that of the Persistence model. This assessment was carried out throughout six brief forecast timeframes. The proposed approach is easier to understand than several studies in the literature, has shown comparable results to more robust deep learning models, and offers a new avenue for future study in forecasting solar photovoltaic energy production.

Miraftabzadeh et.al (2023) [13] proposed a method that use transfer learning to apply known deep learning models from existing photovoltaic (PV) plants to newly constructed PV plants in the same area. The numerical results illustrate the effectiveness of transfer learning in predicting photovoltaic (PV) production one day ahead, especially in newly built PV plants with little historical data. Among the nine models analyzed in this study, the LSTM models exhibit greater performance in forecasting PV power. The use of an inadequate dataset in the novel LSTM model yielded a mean square error (MSE) of 0.55 and a weighted mean absolute percentage error (wMAPE) of 47.07%. Nevertheless, the performance of the LSTM model that was transferred was improved, resulting in a decreased Mean Squared Error (MSE) of 0.168 and a weighted Mean Absolute Percentage Error (wMAPE) of 32.04%.

Seghiour et.al (2023) [14] devised a deep learning (DL) approach to detect, diagnose, and classify the aforementioned errors. The suggested methodology consists of four fundamental stages. Firstly, it employs the Coyote Optimization Algorithm (COA), a heuristic optimization technique, to calculate the values of the five-unknown electrical parameters of the One Diode Model (ODM). These parameters are then integrated into a simulation using the PSIM program, with the objective of faithfully reproducing the behavior of the operational PV system. In addition, the database should be designed to store measurements of current, voltage, and power at the Maximum Power Point (MPP), module temperature, and sun irradiance for the photovoltaic (PV) system. It is necessary to conduct these measures in both healthy and faulty working situations, while providing optimal operational circumstances. The proposed methodology has been validated using data obtained from an operational photovoltaic (PV) facility situated in Algeria. The acquired data have shown the effectiveness of the proposed technique in identifying and categorizing various types of photovoltaic (PV) problems.

Khan et.al (2022) [15] proposed a sophisticated and versatile stacked ensemble method known as DSE-XGB. The system employs two advanced deep learning algorithms, namely artificial neural network (ANN) and long short-term memory (LSTM), as foundational models for predicting solar energy. A method called extreme gradient boosting is used to merge the predictions made by the base models, with the aim of enhancing the accuracy of the solar PV production forecast. The suggested model was evaluated using four separate solar generation datasets to provide a comprehensive and comprehensive assessment. The study used the shapely additive explanation framework to get a comprehensive knowledge of the algorithm's learning process. The effectiveness of the proposed model was evaluated by comparing the predicted results with those of individual ANN, LSTM, and Bagging. The DSE-XGB approach demonstrates exceptional uniformity and reliability in several case studies, irrespective of variations in meteorological conditions. Furthermore, it demonstrates a substantial enhancement in the R2 coefficient, outperforming prior models by a margin of 10%–12%.

SN Venkatesh et.al (2021) [16] presented the expanding photovoltaic industry is preoccupied with the operation and maintenance of photovoltaic (PV) modules. By conducting these evaluations, precise categorization outcomes can be attained with minimal human intervention and time investment. Prompt outcomes are imperative in a dynamic global landscape, where in inventive strategies and technological progress are perpetually evolving. Fault diagnosis is a technique that guarantees the longevity of numerous critical components and generates immediate outcomes. Classification of the derived high-level features from the images is performed using convolutional neural networks (CNN) implemented with the Softmax activation function. A VGG16 network that has undergone prior training is utilized for the purposes of defect classification and feature extraction. The investigation comprises a total of six unique test conditions. Consideration is given to test conditions including burn marks, delamination, discoloration, glass breakage, intact panels, and snail trails. The classification outcome of the pre-trained CNN model is utilized to assess the modeling's performance. In their study, Abubakar et.al (2021) [17] examined the growing use of solar photovoltaics (PV) as an alternative to conventional fossil fuel-based power production, which has seen a significant surge. This increase in interest has prompted research into the advancement of more dependable and efficient maintenance and operating techniques. Regular maintenance is essential for photovoltaic (PV) systems to guarantee consistent generating efficiency. The capacity of artificial intelligence (AI) to replace traditional maintenance methods has gained heightened attention in recent years. The importance of artificial intelligence (AI) in several practical fields, especially in solar photovoltaic applications, is immense. A comprehensive analysis of artificial intelligence approaches used for diagnosing and identifying problems in photovoltaic systems. This study examines several AI-based fault detection and diagnosis strategies proposed in the literature, as well as common defect types that often occur in photovoltaic (PV) systems. Surprisingly, there is a lack of information on the use of PV compared to other industries. In their study, Pierdicca et.al (2020) [18] put up the notion that renewable energy sources are the only feasible alternatives capable of successfully reducing the pollutant emissions linked to the use of fossil fuels. Photovoltaic (PV) power plants are a key technology used to create renewable energy. Monitoring the physiological status of a system is crucial. However, the economic feasibility of using these methods for regular inspections is hindered by their labor-intensive nature, frequent reliance on laboratory equipment, and interference with energy production. Furthermore, photovoltaic (PV) installations are often located in remote locations, making any kind of intervention dangerous. Solar is an AI system that use deep learning to detect irregularities in solar photovoltaic photos taken by unmanned aerial vehicles equipped with thermal infrared sensors. The Mask R-CNN architecture was chosen for the automated inspection task, specifically for the anomalous cell detection system. This architecture, known as the Mask Region-Based Convolutional Neural Network (Mask R-CNN), was selected because it can simultaneously perform object recognition and instance segmentation. The system being examined is trained and assessed using a publicly available dataset of solar thermal pictures that was carefully selected and organized. An evaluation of the effectiveness of three advanced deep neural networks (DNNs) - UNet, FPNet, and LinkNet - via a comparative comparison of their performance. The results provide conclusive proof of the intersection over union (IoU) and Dice coefficient values, confirming the appropriateness and efficacy of the proposed approach.

Good game Kim et.al (2020) [19] examined the need to identify and fix faults in the component elements of photovoltaic (PV) systems to maintain their reliability and safety, as well as to avoid facility accidents and financial losses. The identification of problems in a solar system may be achieved by using the current ratio (IR), power ratio (PR), and voltage ratio (VR) approach. The lower control limit (LCL) and upper control limit (UCL) for each ratio were determined by examining data collected from a functional test site system. The algorithm identified the PR as incorrect when it exceeded the defined range. The approach was then used to categorize system errors as either total, parallel, or serial for PR and IR. The results showed that the PR values were outside the expected range of 0.93 to 1.02 for series, total, and parallel faults. Specifically, the PR values for series faults ranged from 0.91 to 0.68, for total faults ranged from 0.88 to 0.62, and for parallel faults ranged from 0.66 to 0.33. Moreover, in the presence of series and total faults, VR demonstrated superior performance compared to the LCL (0.99) and UCL (1.01) by a margin of 0.95-0.69 and 0.91-0.62, respectively. However, this discrepancy was not detected in the presence of parallel faults.

Zyout et.al (2020) [20] investigated the growing importance of automatic solar panel flaw inspection due to the rise in worldwide production and installation of solar energy systems. Deep convolutional neural networks (CNNs) provide outstanding results when used for image categorization in many applications. The problem has been detected and the characteristics of the PV panel's surface are characterized using convolutional neural networks. The method's combination of transfer learning and AlexNet CNN yielded very promising results in detecting various surface flaws on solar panels. There is a wide range of authors who studied on the solar PV system using deep learning approaches and give their findings as shown in Table 2.2.

Table 2.2 Comparison of reviewed literature

Authors & Year	**Methods**	**Key Findings**	**Future Work**
Kumar et.al (2024) [11]	Hybrid deep learning (LSTM-RNN)	The LSTM model surpasses other models in predicting solar power, as indicated by its superior error metrics: a mean	Enhance prediction accuracy and operational efficiency

		squared error (MSE) of 0.0091.	
Maciel et.al (2024) [12]	ANN and GBM	Integrates Artificial Neural Network and Light Gradient Boosting Machine, showed an average forecast accuracy that was 17.5% superior than that of the Persistence model	Forecasting solar photovoltaic energy production
Miraftabzadeh et.al (2023) [13]	LSTM	The LSTM model that was transferred was improved, resulting in a decreased Mean Squared Error (MSE) of 0.168	Focus on enhancing the model's robustness by incorporating diverse weather data from multiple geographical locations
Seghiour et.al (2023) [14]	COA and ODM	The acquired data have shown the effectiveness of the proposed technique in identifying and categorizing various types of photovoltaic (PV) problems.	Focus on enhancing the scalability of models to handle diverse environmental conditions and various fault types.
Khan et.al (2021) [15]	DSE-XGB	It demonstrates a substantial enhancement in the R2 coefficient, outperforming prior models by a margin of 10%–12%.	Explore incorporating real-time weather data

SN Venkatesh et.al (2021) [16]	VGG16 CNN with Softmax activation	VGG16 CNN accurately classifies six unique test conditions for PVM flaws. Consideration of various defect types.	Continued assessment of modeling performance with different defect categories. Exploration of other CNN architectures for defect identification.
A Abubakar et.al (2021) [17]	AI-based fault detection and diagnosis techniques for PV systems	Examination of various AI-based fault detection techniques in PV systems. Scarce information in literature regarding PV applications.	Further research on AI techniques for fault detection in PV systems. Exploration of fault types specific to PV applications.
R Pierdicca et.al (2020) [18]	Anomalous cell detection using Mask R-CNN architecture, comparative analysis with UNet, FPNet, and LinkNet	Mask R-CNN architecture effectively identifies anomalies in solar thermal images. Comparative analysis with three DNNs demonstrates suitability and effectiveness.	Continued research on DNN architectures for automated inspection tasks. Exploration of other cutting-edge DNNs for comparison.
GG Kim et.al (2020) [19]	Fault identification using current ratio (IR), power ratio (PR), and voltage ratio (VR)	PR and VR ratios used for fault identification in PV systems. Algorithm classified system faults as total, parallel, or series.	Further refinement of algorithms for fault identification. Exploration of additional ratios for improved fault classification.
I Zyout et.al (2020) [20]	Deep CNN for automatic solar panel defect inspection	Deep CNN, specifically AlexNet, exhibits exceptional performance in detecting surface defects on solar panels.	Exploration of other deep CNN architectures for defect inspection. Further enhancement of transfer

			learning approaches for defect identification.

2.4 Research Gap

In reviewing the literature surrounding the predictive maintenance of photovoltaic (PV) power installations, several significant gaps become evident, particularly concerning the integration and optimization of fault detection systems. While current studies have introduced various methodologies, including deep learning techniques, to improve fault detection in solar panels, these approaches often operate in isolation rather than as part of a cohesive system. The literature reveals four primary research gaps:

1. **Integration of Multi-modal Data Sources**: Existing research predominantly focuses on singular methodologies, either leveraging image data through convolutional neural networks (CNNs) and analyzing operational data. There is a notable deficiency in the synthesis of these diverse data types. A comprehensive model that effectively integrates both image and tabular data to leverage the complementary strengths of different sensory inputs remains underexplored. This integrated approach could potentially offer more robust fault detection capabilities by providing a multi-faceted view of panel health.
2. **Absence of PV Panel Fault Detection Model:** Despite the growing importance of maintaining the efficiency and longevity of photovoltaic (PV) systems, there remains a notable gap in the availability of specialized, pre-trained models dedicated to detecting faults in solar panels directly out of the box. The absence of such models compels researchers and engineers to rely on generic deep learning frameworks. These require significant adaptation and fine-tuning with domain-specific data to achieve the necessary precision for effective fault detection in solar panels. This gap highlights a critical area for further research and development, suggesting a strong potential for advancements in specialized diagnostic tools that can enhance predictive maintenance practices for solar energy systems.
3. **Real-time Fault Detection and Predictive Analytics**: While many models excel in retrospective fault analysis, there is a crucial need for real-time detection systems that predict faults before they lead to system inefficiencies. Current models often do not address the real-

time application of predictive maintenance technologies, which is essential for minimizing downtime and operational losses in solar power installations.

4. **Scalability and Adaptability in Diverse Environmental Conditions**: The effectiveness of fault detection models across various geographical and climatic conditions is not sufficiently explored. Most studies do not account for the adaptability of the models to different environmental factors that significantly influence solar panel performance, such as partial shading, diverse weather patterns like sandstorm which leads to accumulation of dust, and varying degrees of soiling. There is a need for adaptive models that can recalibrate based on localized data inputs, ensuring consistent performance regardless of external conditions.

Addressing these gaps would not only enhance the reliability and efficiency of fault detection systems but also improve the overall energy output and longevity of solar power installations. Future research should focus on developing integrated, real-time, and environmentally adaptive predictive maintenance frameworks that support the dynamic needs of global solar energy infrastructure. This will ensure the sustainability and economic viability of solar power as a key component of the global energy mix.

CHAPTER 3

PROBLEM EXPLANATION

3.1 Background of the Problem

The rise of photovoltaic (PV) power installations has been substantial in recent years, reflecting a global shift towards renewable energy sources. PV systems play a crucial role in sustainable power generation, but their efficiency and longevity are heavily dependent on the condition of the solar panels. Various faults such as microcracks, soiling, delamination, physical damage, and electrical defects can significantly reduce the energy output and increase maintenance costs. Traditional methods of fault detection, which rely on manual inspections, are time-consuming, labour-intensive, and prone to human error, making it difficult to detect and address faults promptly. This inadequacy necessitates the development of advanced fault detection methods that can provide timely and accurate maintenance interventions.

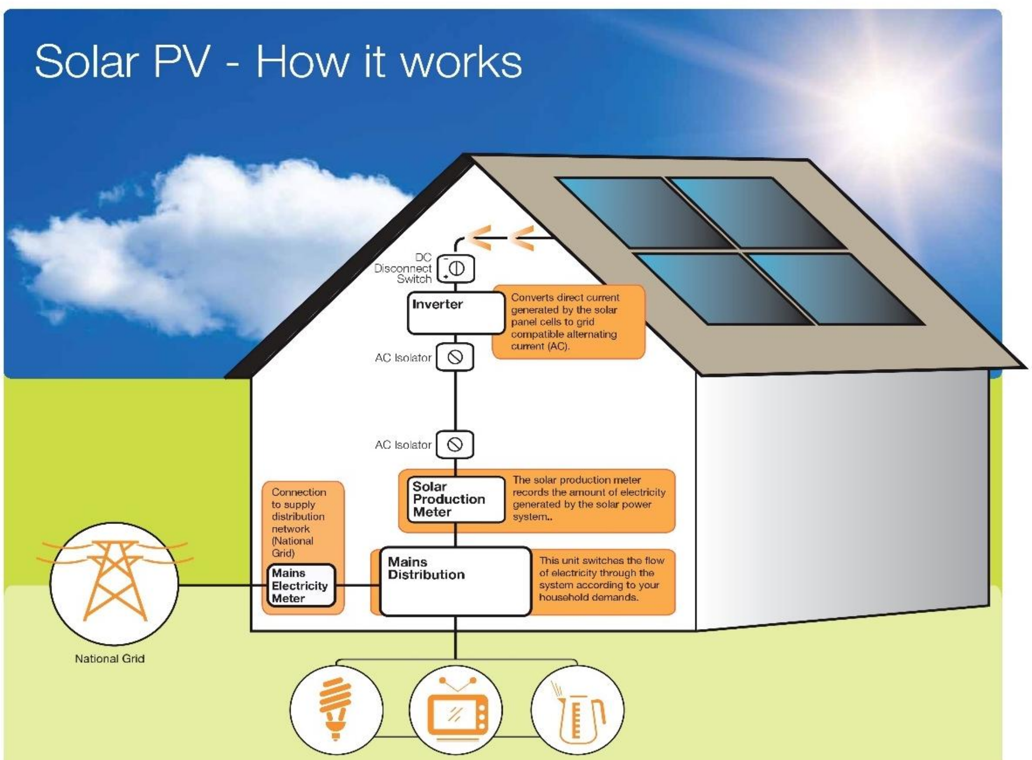


Figure 3.1: Typical setup of a power system for a PV panel power supply system

3.2 Objective

The primary objective of this research is to enhance the predictive maintenance of PV panels by integrating advanced image analytics into existing solar power forecasting models. This involves:

- **Develop a Model**: Construct and validate a predictive maintenance model that leverages both image and tabular fault information to detect and classify panel faults & electrical faults more accurately.
- **Enhance Fault Detection Accuracy**: Implement a novel architecture to enhance the system's ability to identify various fault types under diverse operational conditions.
- **Evaluate Model Performance**: Assess and compare the efficacy of the novel model against traditional monitoring methods and models using single data sources in terms of various evaluation metrices.
- **Fault Area Measurement**: Perform segmentation of the PV panel fault and try to measure the area of the segmented faults using Shoelace formulation. This mathematical approach helps in quantifying the size of the faults detected in the PV panels, providing an additional metric for assessing the severity of the faults.

By achieving these objectives, this research aims to improve the overall performance and longevity of PV systems, contributing to more reliable and efficient solar power generation.

3.3 Software Workflow Setup

The visual representation of our PV Panel Image & Video Capturing System setup is illustrated in Figure 3.2. The system comprises primary and secondary cameras, drones, and satellite captures, which collect comprehensive visual data of photovoltaic panels. The video capturing system involves framewise extraction from video feeds, while the image capturing system collects RGB images. This data is crucial for the segmentation, identification, and planning algorithms employed to improve fault detection and maintenance processes.

The core activities within our research scope, highlighted in green, involve image processing to detect, classify, and segment faults, and measure their areas. These steps directly contribute to the development of a predictive maintenance model aimed at enhancing the reliability and efficiency of solar power generation. Out of scope activities, highlighted in red, include structural integrity assessment, PV panel material type consideration, thermal anomaly detection, and panel alignment analysis.

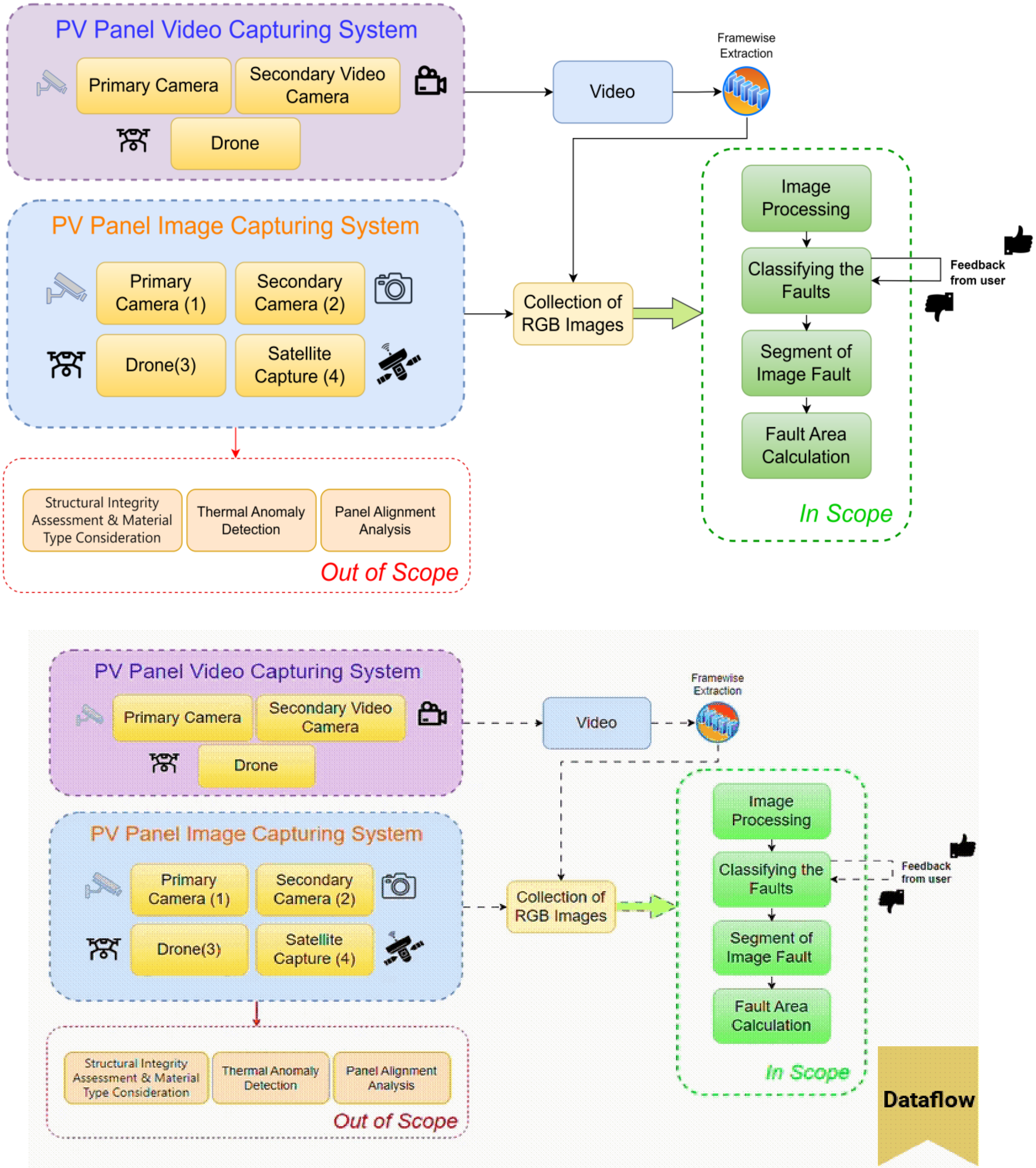


Figure 3.2: Software Workflow for End-to-End Fault Detection and Segmentation

3.4 Tasks Relevant to the Objective

To achieve the stated objectives, the following tasks are outlined:

1. **Data Collection:** We are aggregating a comprehensive dataset that includes images of both faulty and non-faulty photovoltaic (PV) panels along with corresponding tabular data detailing operational parameters. Initial efforts have successfully compiled instances encompassing various fault types and normal operational states to ensure a robust dataset for analysis.
2. **Frame Extraction and Analysis:** Development is ongoing for an algorithm capable of extracting individual frames from video data and isolating PV panel images. This process optimizes the clarity and relevance of visual data extracted from general video footage.
3. **Data Preprocessing:** Preprocessing pipelines for both tabular and image data are established, including imputation of missing values, normalization, and scaling for tabular data, as well as resizing, grayscale conversion, noise reduction, rotation, cropping, and data augmentation for image data. Preliminary tests confirm these enhancements significantly improve data quality and readiness for analytical processing.
4. **Image & Tabular Faults Classification:** We are testing different machine learning models to create an optimized algorithm that can classify images of photovoltaic (PV) panels as either faulty or non-faulty. Additionally, the algorithm aims to identify different types of tabular faults based on various combinations of current and voltage parameters. The initial results are promising, showing that the models can effectively distinguish between the defined classes.
5. **Incremental Training:** To enhance the accuracy of our initial trained model and ensure it can handle a diverse range of data, we have developed a feedback mechanism that allows user input on the classification results. This feedback is then used to retrain the model incrementally. This iterative process not only improves the model's accuracy over time but also enhances its robustness in handling a wide variety of data.
6. **Fault Annotation:** We have drafted guidelines for manually annotating images to identify and delineate areas affected by faults. A subset of images has been annotated to train preliminary segmentation models, facilitating automated fault localization in subsequent tasks.
7. **Image Segmentation:** Our research is currently focusing on experimenting with several advanced image segmentation models, including U-Net and Mask R-CNN. The goal is to

enhance the granularity and accuracy of fault detection by precisely highlighting and isolating faulty regions within the PV panels.

8. **Area Calculation:** The integration of area calculation algorithms using the Shoelace formula within the segmentation process is being refined. This step will provide accurate measurements of the fault areas, which is crucial for assessing the extent of damage and the overall integrity of the panels.

3.5 Formulation of the Problem

In our current study, we address two principal challenges concerning the maintenance and efficiency optimization of photovoltaic (PV) systems. The first challenge involves the classification of solar panel faults using both image and tabular data. The second challenge is the quantification of these faults through the calculation of the fault area as a percentage of the total panel area.

3.5.1 Image Fault Classification:

Image classification in the context of PV panels involves the categorization of pixel data into predefined classes that represent different types of faults. This process is akin to mapping numerical values of pixel intensity to symbolic categories that signify various fault conditions such as cracks, soiling, or delamination.

Mathematical Formulation:

Let X be a pixel vector in an image, and $D = \{c_1, c_2, \dots, c_L\}$ be the set of fault categories. The classification function can be expressed as:

$$f(x): X \rightarrow D$$

where each pixel vector X is assigned to one fault category in D.

3.5.2 Image Segmentation and Percentage of Fault Area Discovery:

Following classification, image segmentation is employed to delineate and quantify the exact areas affected by faults. This step is critical for calculating the percentage of the panel's area that is compromised, which directly influences the panel's performance and longevity.

Mathematical Formulation:

If S denotes the segmented area of faults and A represents the total area of the solar panel, the percentage of the fault area is given by:

Percentage of Fault Area = $\left(\frac{S}{A}\right) \times 100\%$

Using the Shoelace formula (Gauss's area formula for simple polygons), the area of a fault S segmented in an image can be calculated if the vertices of the polygon are known:

$$S = \frac{1}{2}\left|\sum_{i=1}^{n-1} (x_i y_{i+1} - y_i x_{i+1}) + (x_n y_1 - y_n x_1)\right|$$

3.5.3 Fault Classification in Tabular Data

In addition to image data, we also explored and collected tabular fault data. Based on the parameters of current and voltage, three categories of faults were identified and measured.

Features and Labels

- Let $X = [x_1, x_2, \dots, x_n]$ be the feature matrix where each x_i represents an individual data point with features such as currents, voltages, power, etc.
- Let $y = [y_1, y_2, \dots, y_n]$ be the vector of labels corresponding to each data point in X, where y_i corresponds to the fault classes.

Mathematical Formulation:

Objective Function

- The loss function for logistic regression via SGD can be defined as:

$$L(y, f(x; \theta)) = -y\log(p) - (1 - y)\log(1 - p)$$

where $p = P(y = 1 \mid x; \theta)$ is the predicted probability from logistic regression.

Parameter Update

- Parameters are updated incrementally for each training example:

$$\theta \leftarrow \theta - \eta \nabla_\theta L$$

where η is the learning rate, and $\nabla_{\theta} L$ is the gradient of the loss function evaluated at the current example.

Training and Prediction

- Choosen model is trained using the dataset (X, y), adjusting θ to minimize the respective loss functions.
- The trained model can then classify new data points, predicting whether they correspond to a fault class based on their features.

3.6 Details Explanation

3.6.1 Constraints

The problem formulation includes several constraints to ensure practical and effective fault detection and area estimation:

- Physical Constraints: $A \geq 0$, as fault area cannot be negative.
- System Constraints: $A \leq A_{\max}$, where $A_{\max}$ is the maximum area of the solar panel.
- Tabular Data: Voltage and current parameters cannot be negative.

The goal is to classify and segment faults in PV panels using both images and tabular fault data and accurately measure the fault area.

3.6.2 Formulation:

- Classification(both image and tabular fault) and Segmentation: $F1, F2, \ldots, Fn$
- Fault Area Measurement: $A1, A2, \ldots, An$
- Model: f with parameters θ, predicting $\hat{F} = f(X; \theta)$ and $\hat{A} = g(X; \theta)$

3.6.3 Optimization Goal:

$$\min_{\theta} \sum_{t-1}^{T} \left(L\big(F, f(X; \theta)\big) + L(y, f(x; \theta)) + \alpha \times E(A), g(X; \theta)) - \lambda \times (1 - R)\right)$$

The learning approach involves using a combination of image fault & tabular fault classification, segmentation, and area measurement to train models for fault detection and quantification. Below is the term wise explanation,

$\boldsymbol{L(F, f(X;\theta))}$ **:**

- **Purpose**: This term represents the loss function for image classification.
- **Explanation**: F denotes the true fault categories from image data, and $f(X;\theta)$ represents the predictions made by the image classification model parameterized by θ. This component of the loss measures how well the model predicts image classes (e.g., types of faults in PV panels).

$\boldsymbol{L(y, f(x;\theta))}$ **:**

- **Purpose**: This term corresponds to the loss for tabular data classification.
- **Explanation**: Here, y is the actual class labels from the tabular data, and $f(x;\theta)$ signifies the predictions from the tabular data classification model. This loss assesses the accuracy of fault predictions based on operational data like voltage and temperature readings.

$\boldsymbol{\alpha \times E(A, g(X;\theta))}$ **:**

- **Purpose**: This term is the loss associated with image segmentation.
- **Explanation**: A indicates the true segmented areas (e.g., actual areas of faults), while $g(X;\theta)$ is the output from the segmentation model. E quantifies the error in segmenting the image, essentially how accurately the model can delineate fault areas within the image.

$\boldsymbol{\lambda \times (1 - R)}$ **:**

- **Purpose**: This term is a penalty that increases as the recall R decreases.
- **Explanation**: Recall R is critical in fault detection systems as it measures the proportion of actual positives that are correctly identified. High recall is crucial in applications where failing to detect a fault (false negative) can have serious implications. λ is a weighting factor that adjusts the influence of recall on the overall loss.

3.6.4 Objective and Optimal Parameter:

The objective is to minimize classification and area estimation errors while ensuring high recall, thus accurately detecting and quantifying faults and minimizing false negatives.

Optimization

The optimization process involves finding the optimal parameters θ^* by minimizing the loss functions L and E, and maximizing recall.

Optimal Parameter:

$$\theta^* = \arg\ \min_{\theta} \frac{1}{N}\sum_{i-1}^{N} \left(L\big(y_i, f(X_i;\theta)\big) + L(y, f(x;\theta)) + \alpha \times E\big(a_i, g(X_i;\theta)\big) - \lambda \times (1-R)\right)$$

Significance of Recall value in penalty term $\boldsymbol{\lambda} \times (\mathbf{1} - \boldsymbol{R})$:

In the context of photovoltaic (PV) panel fault detection and quantification importance of recall (R) is as follows,

- **High Stakes**: Missing a fault (a false negative) could mean not addressing a panel that is underperforming or at risk of further damage. This could reduce the overall efficiency of the solar power system and lead to higher maintenance costs or system failure.
- **Safety**: Undetected faults might lead to safety hazards, depending on the nature of the fault (e.g., electrical issues leading to fire risks).

3.7 Integrating Feedback for Image and Tabular Data Classification

This study incorporates user feedback mechanisms to refine both image and tabular data classification accuracy and corresponding mathematical formulation will be explained here.

3.7.1 Image Data Feedback Integration:

Image data if it's not classified correctly then user give feedback to that image with its correct class and our model will retrain it (partially fit) with only new set of categorized images to enhance the model accuracy.

Notation:

- X : Input image data.
- Y: True labels for the data.
- $\hat{Y}$: Predicted labels
- θ : Parameters

Feedback Mechanism:

- Upon receiving user feedback indicating incorrect classifications, the specific data entries are flagged.
- These entries are then added to a retraining batch with corrected labels to update the model.

Loss Function Update:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} L\big(y_i, f(x_i; \theta)\big)$$

Here, L is a loss function, such as cross-entropy, measuring the difference between predicted and actual labels.

Model Update:

The trained model parameters θ are updated by minimizing the loss function using gradient descent:

$$\theta = \theta - \eta \nabla_\theta L(\theta)$$

$where\ \eta$: Learning rate.

3.7.2 Tabular Data Feedback Integration:

For tabular fault data, feedback is integrated into the stochastic gradient descent (SGD) logistic regression model to categorize faults more accurately.

Notation:

- X : Feature matrix from tabular data.

- Y : True fault categories.
- $\hat{Y}$: Predicted fault categories.

Feedback Mechanism:

Similar to image data, feedback on misclassifications is used to adjust the labels in the training dataset.

Loss Function Update:

$$L(y, f(x;\theta)) = -y\log(p) - (1-y)\log(1-p)$$

where $p = P(y = 1 \mid x;\theta)$ is the predicted probability from logistic regression

Model Update:

Update the parameters θ by applying SGD:

$$\theta = \theta - \eta\nabla_{\theta}L(\theta)$$

Objective of Feedback Integration:

- **Enhance Model Accuracy**: By continually integrating feedback, the models adapt to new data and corrections, improving their predictive performance over time.
- **Reduce Misclassifications**: Targeted retraining on misclassified instances helps in reducing overall error rates in fault classification.

3.8 Out of Scope

Below are the details out of scope of the current research,

1. **Material Type Consideration**: The exploration of different types of photovoltaic (PV) panel materials such as monocrystalline, polycrystalline, and amorphous silicon is recognized as an important aspect of photovoltaic research.
2. **Fault Type Limitation**: The current research is limited to identifying and classifying specific types of faults on PV panels, including dust accumulation, bird droppings, snow cover, electrical faults, and physical damages.

3. **Panel Alignment Analysis**: This study does not essentially involve accessing the tilt and orientation of the panels relative to the sun's position throughout the year.
4. **Thermal Anomaly and Structural Integrity Analysis**: Finally, this analysis does not focus on detecting thermal anomalies and assessing the structural integrity of PV panels.

3.9 Challenges

- One significant challenge we encounter is related to the computational resources required for my research. we primarily utilize Google Colab for model development and training, which provides access to GPU resources. However, we frequently reach the GPU usage limit by mid-day. Once this limit is exhausted, we are restricted to using only the available CPU resources, which are considerably less efficient for processing deep learning tasks. This constraint significantly slows down my progress, as the computational power of CPUs does not suffice for the intensive calculations that my models require, leading to prolonged training times and delayed experimentation cycles.
- We anticipate challenges related to the variability of image quality and the subtle nature of some faults, which may complicate the training process. Additionally, integrating these image analytics into our existing predictive maintenance framework will require careful calibration to ensure seamless operation and timely fault detection.
- The manual annotation of each fault in the images is a particularly daunting task, requiring meticulous attention to detail. This process demands a significant time investment, as each image must be carefully inspected and labeled to ensure accurate and reliable data for training the deep learning models. Additionally, variations in image depth pose further challenges, making it essential to maintain consistency and accuracy in the annotation process to effectively train the models.

CHAPTER 4

DATASETS

4.1 Overview

In the development of effective AI models, the utilization of high-quality datasets is fundamental. Given the objectives of our study, which involve the detection of PV panel faults and identifying the fault area along with percentage of damage, a supervised approach necessitating labeled data is essential. This section provides a detailed description of the datasets utilized in our study, including their sources, structures, and intended purposes.

4.1.1 Dataset Description

Link:

1. https://www.kaggle.com/datasets/pythonafroz/solar-panel-images
2. https://www.kaggle.com/datasets/amrezzeldinrashed/fault-detection-dataset-in-photovoltaic-farms
3. https://www.kaggle.com/datasets/marcosgabriel/photovoltaic-system-thermography
4. https://www.pexels.com/search/videos/solar%20panel/

Explanation:

Panel Images Dataset:

- **Solar Source**: Kaggle (uploaded by user pythonafroz).
- **Description**: The accumulation of contaminants such as dust, snow, and bird droppings on the surfaces of solar panels significantly impairs the efficiency of these modules, leading to a reduction in energy output. Thus, it is critical to develop and refine monitoring and cleaning methodologies for solar panels to enhance their efficiency, minimize maintenance costs, and reduce resource consumption.

  The aim of this dataset is to evaluate the efficacy of various machine learning classifiers in detecting the presence of dust, snow, bird droppings, as well as physical and electrical damage on solar panel surfaces, with the goal of achieving the highest possible accuracy.

Directory Overview: This directory organizes the data into six distinct class folders, each representing a different condition of the solar panels, facilitating classification tasks:

- **Clean:** Contains images of solar panels without any visible faults.
- **Dusty**: Includes images showing solar panels with dust accumulation.
- **Bird-drop**: Features images with bird droppings on solar panels.
- **Electrical Damage**: Comprises images displaying solar panels with electrical damage.
- **Physical-Damage**: Contains images of solar panels with physical damage.
- **Snow-Covered**: Includes images of solar panels covered in snow.

- **Purpose**: This dataset could be used for various purposes, including but not limited to:
  - **Solar panel detection and segmentation**: Developing algorithms to automatically detect and segment solar panels within images.
  - **Solar panel condition monitoring**: Analyzing the condition of solar panels based on their appearance in images, such as detecting cracks, soiling, or other forms of damage.
  - **Solar panel performance analysis**: Studying the performance of solar panels based on their visual appearance, such as identifying areas with reduced efficiency due to shading.

Sample faulty panel image looks like below,

Table 4.1: Different PV Panel Fault

Electrical Faults | Bird Drops | Dusty Panel | Physical Damage | Snow Cover

Below are the class wise counts,

Table 4.2: Fault wise count of PV Panel Image

Bird Drop	***Clean***	***Dusty***	***Electrical Damage***	***Physical Damage***	***Snow***
178	193	190	103	69	123

Table 4.3: Description of each fault type

Class	***Description***
Clean	Contains images of solar panels without any visible faults. This category is used as a baseline for comparison against faulty panels.
Dusty	Includes images showing solar panels with dust accumulation. Dust can significantly reduce panel efficiency by blocking sunlight.
Bird-drop	Features images with bird droppings on solar panels. Such obstructions can cause hot spots and reduce panel performance.
Electrical Damage	Comprises images displaying solar panels with electrical damage. This can include burnt marks or discoloured areas indicating electrical issues.
Physical-Damage	Contains images of solar panels with physical damage such as cracks or shattered glass. Physical damage can lead to significant energy losses and safety hazards.
Snow-Covered	Includes images of solar panels covered in snow. Snow coverage can completely block sunlight, halting energy production.

Apart from the above faulty data we have also collected some more additional PV panel faults image with one new additional class called “Red Dust”.

Table 4.4: Different PV Panel Fault for additional dataset

Below are the class wise counts for additional dataset,

Table 4.5: Fault wise count of PV Panel Image for additional dataset

Bird Drop	***Clean***	***Dust***	***Red Dust***	***Snow***
300	551	281	220	264

Table 4.6: Description of additional fault type in additional data

Class	***Description***
Red Dust	Contains images of solar panels showing red dust accumulation which educe the panel efficiency by blocking the sunlight.

<u>Panel Video Dataset:</u>

- **Source**: Pexels
- **Description:** This website contains huge collection of videos. We have search for Solar Panel Videos and use Pexels API_KEY to extract those videos programmatically and stored for our further study.

Sample faulty panel videos looks like below,

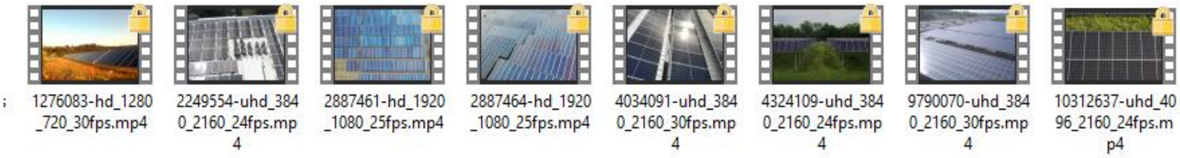
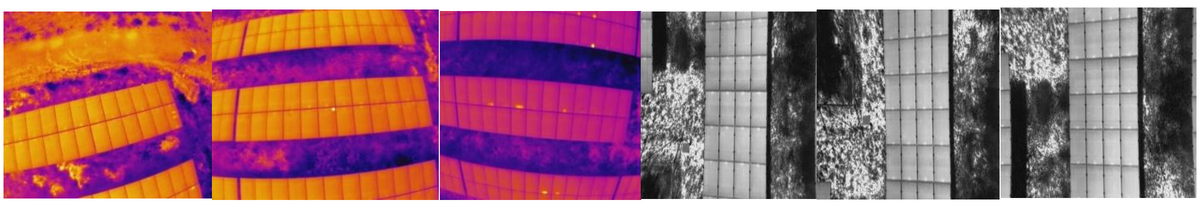


Figure 4.1: Sample Videos of Solar Panel and corresponding Faults

<u>Faulty PV Panel Thermal & Infrared Images:</u>

- **Solar Source**: Kaggle (uploaded by user MARCOS GABRIEL).

Table 4.7: Different Faulty PV Panel Thermal& Infrared Images



<u>Tabular Fault Detection Dataset in Photovoltaic Farms:</u>

- **Source**: Kaggle (uploaded by user amrezzeldinrashed).
- **Description**: For the purpose of this study, a 250-kW simulated photovoltaic (PV) power plant was employed to generate both training and testing datasets relevant to various PV fault scenarios. Within this framework, three distinct types of faults along with a normal operational state (free of faults) were delineated. According to the illustrative data provided, the fault types, labelled F1, F2, and F3, correspond to a string fault (examined on string 1), a string-to-ground fault (also on string 1), and a string-to-string fault (assessed between strings 1 and 2), respectively. The constructed datasets comprise a training set with 600 instances, each characterized by 30 features and a designated class or category column.
    - **Training Dataset:** It includes 100 instances (16.67%) representing the free-of-fault state, 153 instances (25.5%) with string faults, 149 instances (24.83%) with string-to-ground faults, and 198 instances (33%) with string-to-string faults. The simulations were conducted over a total duration of 0.4 seconds, assuming the occurrence of faults at the midpoint, i.e., 0.2 seconds. All data measurements in the training set were recorded post-fault, between 0.2 and 0.4 seconds.

- **Testing Dataset:** It encompasses 50 instances, with data collection spanning from 0.1 to 0.3 seconds. This includes a transient phase from 0.1 to 0.2 seconds before faults were simulated to occur from 0.2 to 0.3 seconds. This temporal structure was designed to rigorously test the model's responsiveness and accuracy in fault detection under dynamically changing conditions.

- **Purpose**: This dataset is likely aimed at detecting faults or anomalies in PV systems, which could include various issues such as:
  - **Performance degradation**: Identifying factors leading to reduced energy output from solar panels, such as shading, soiling, or degradation over time.
  - **Equipment malfunction:** Detecting faults in components of the PV system, such as inverters, connectors, or individual solar panels.
  - **Environmental factors**: Analyzing the impact of environmental conditions like temperature, humidity, and solar irradiance on the performance and health of the PV system.

Table 4.8: Field information of fault Detection Dataset in Photovoltaic Farms

Attribute	***Description***
Source	Kaggle (uploaded by user amrezzeldinrashed)
Training Dataset	600 instances characterized by 30 features and a designated class or category column
Fault Types	F1 (string fault), F2 (string-to-ground fault), F3 (string-to-string fault)
Normal State	Free-of-fault state (100 instances, 16.67%)
String Faults	153 instances (25.5%)
String-to-Ground Faults	149 instances (24.83%)
String-to-String Faults	198 instances (33%)
Testing Dataset	50 instances collected over 0.2 seconds, testing model responsiveness and accuracy in fault detection

CHAPTER 5

RESEARCH AND METHODOLOGY

5.1 Overview

Building upon the research objectives outlined in Chapter 1 and addressing the research gaps identified in the literature review (Chapter 2), this chapter details the methodology used to investigate and solve the problem of photovoltaic (PV) panel fault classification and segmentation. The methodology is designed to achieve the objectives of developing a predictive maintenance model, enhancing fault detection accuracy, and quantifying fault areas, as discussed in Chapter 3.

This chapter details the research design and methodologies employed to develop a novel image classification algorithm integrating Dynamic PCA and a Hierarchical Ensemble Classifier. This section also outlines the approach for extending the classification framework to include image segmentation for fault detection and area measurement in photovoltaic (PV) panels.

5.2 Algorithm Overview:

To address the need for accurate and efficient fault detection in PV panels, as highlighted in the problem explanation (Chapter 3), this study employs a combination of advanced image analysis and machine learning techniques. The methodology is structured to integrate both image and tabular data, ensuring a comprehensive approach to fault detection.

5.2.1 Adaptive Dimensionality Reduction and Classifier Fusion for Image Fault Classification

Incremental PCA for Image Classification

Incremental Principal Component Analysis (IPCA) is an adaptive dimensionality reduction technique that optimizes the number of principal components for different subsets of data, rather than applying a fixed number globally. This method preserves essential features by capturing local data variations, which is particularly beneficial in image classification where dataset heterogeneity is common. By efficiently processing data in mini-batches, IPCA enhances classifier performance while maintaining computational efficiency.

SVM Classifier for Image Classification

Support Vector Machine (SVM) classifiers are powerful tools for image classification due to their ability to handle high-dimensional data and capture complex, non-linear relationships through kernel functions. SVMs work by finding the optimal hyperplane that maximally separates different classes in the feature space, thus providing robust initial classification. Their effectiveness in managing both linear and non-linear data structures makes SVMs particularly suitable for diverse and intricate image datasets.

Gaussian Naive Bayes for Image Classification

Gaussian Naive Bayes (GNB) is a probabilistic classifier that applies Bayes' theorem with the assumption of feature independence and a Gaussian distribution of continuous variables. In image classification, GNB is valued for its simplicity and computational efficiency. By focusing on the statistical properties of pixel intensities, GNB effectively classifies images, especially when integrated with feature scaling techniques that emphasize the importance of key features identified by prior dimensionality reduction methods like PCA.

5.2.2 Enhanced Logistic Regression Classifier for Tabular Fault Classifier

Logistic Regression and Stochastic Gradient Descent (SGD)

Logistic Regression and Stochastic Gradient Descent (SGD) Classifier are robust machine learning methods ideal for classifying tabular fault data in predictive maintenance systems, like those used for monitoring photovoltaic panels. Logistic Regression provides a probabilistic framework, making it effective for predicting the likelihood of specific fault types based on various operational metrics such as voltage, current, and temperature. On the other hand, the SGD Classifier, due to its efficiency in handling large datasets, is suitable for optimizing the logistic regression model, especially when the data is vast or streamed in real-time. By employing SGD, the logistic regression model can be iteratively updated, ensuring quick adaptation to new data and maintaining high performance in dynamic operational environments. Together, these methods form a powerful duo for fault classification, leveraging logistic regression's predictive clarity and SGD's computational efficiency.

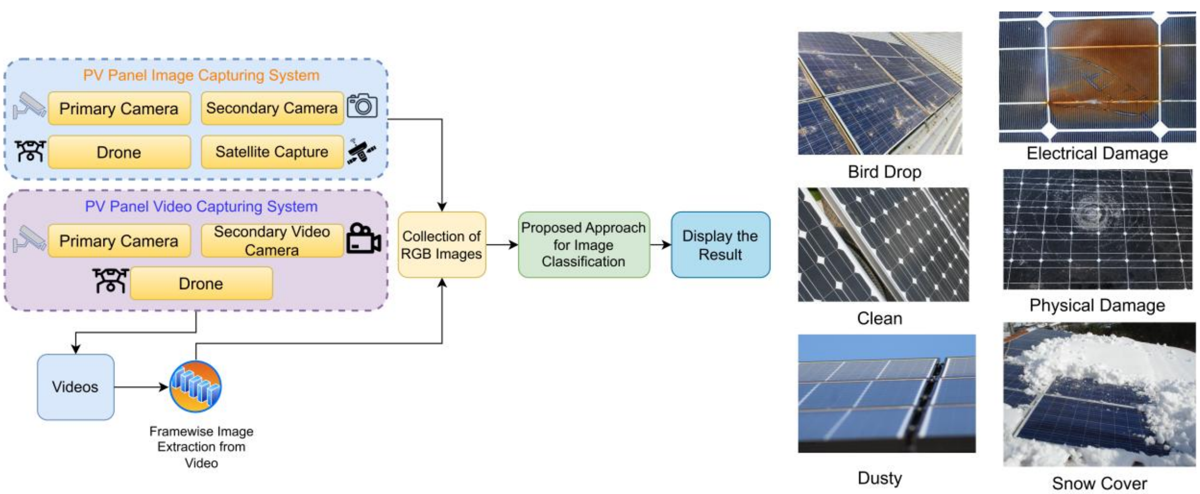
5.3 PV Panel Identification & Panel Fault Classification using Images & Videos

The first part of methodology will explain the novel approach for PV panel identification and panel fault classification technique used in this study to identify the PV panel and further classify the panel faults from the RGB image and videos.

5.3.1 Integrating Adaptive Feature Selection with Hierarchical Ensemble Methods for Robust Image Classification – A Novel Approach for PV Panel Identification & Panel Fault Classification Using Incremental PCA and Hierarchical Ensemble Methods

In the field of image classification, traditional methods such as Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) have demonstrated significant success. However, there exists a compelling need for approaches that offer not only high accuracy but also computational efficiency and adaptability to varying data characteristics. This section proposes a novel methodology that combines Dynamic Principal Component Analysis (PCA) with a Hierarchical Ensemble Classifier. This approach aims to enhance classification performance by dynamically reducing dimensionality, improving decision boundaries, and integrating feature importance into classification models.

Fig. 5.1 & 5.2 shows the detailed proposed architecture.



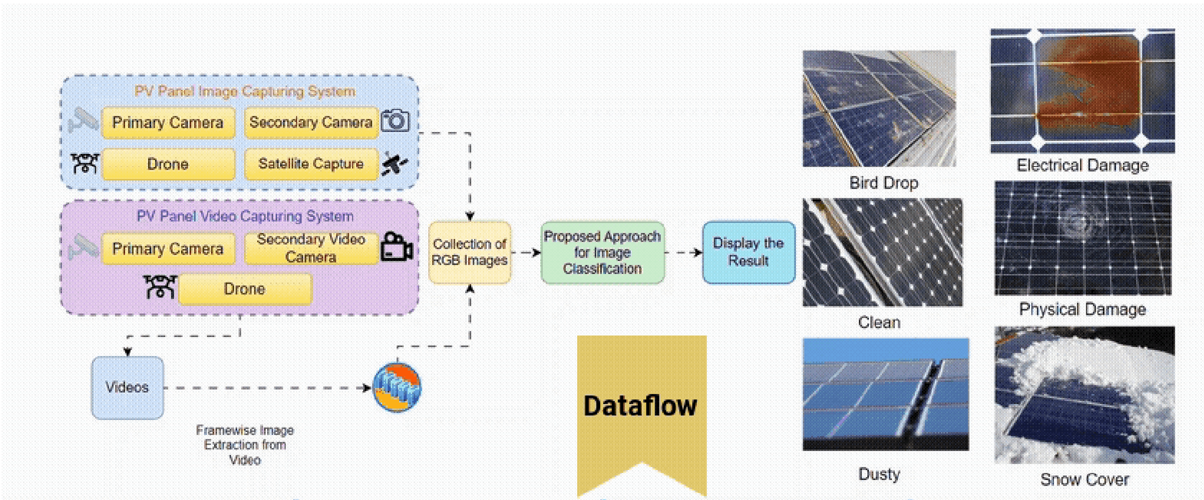
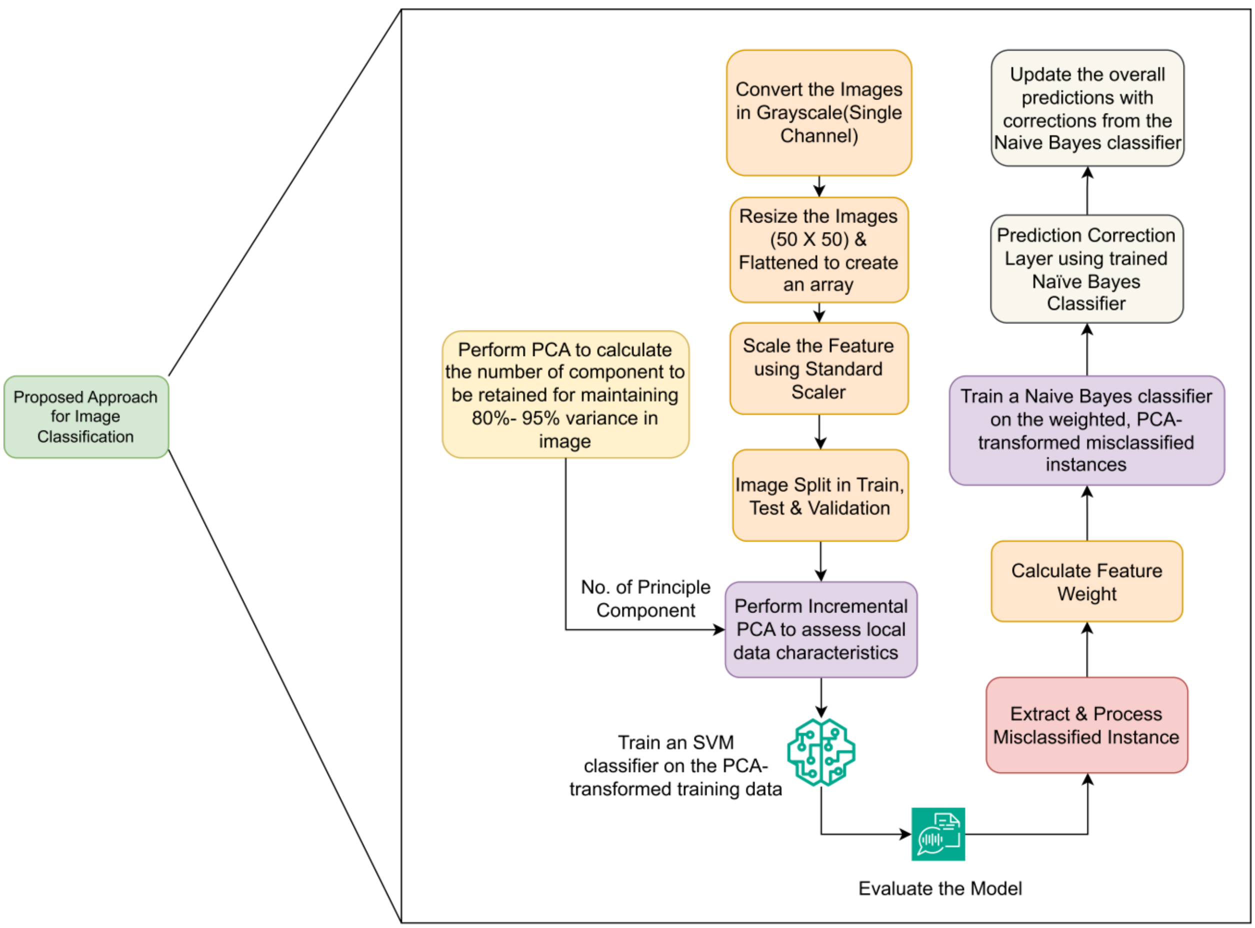


Figure 5.1: Proposed Framework for Faults Classification from PV Panel Images & Videos



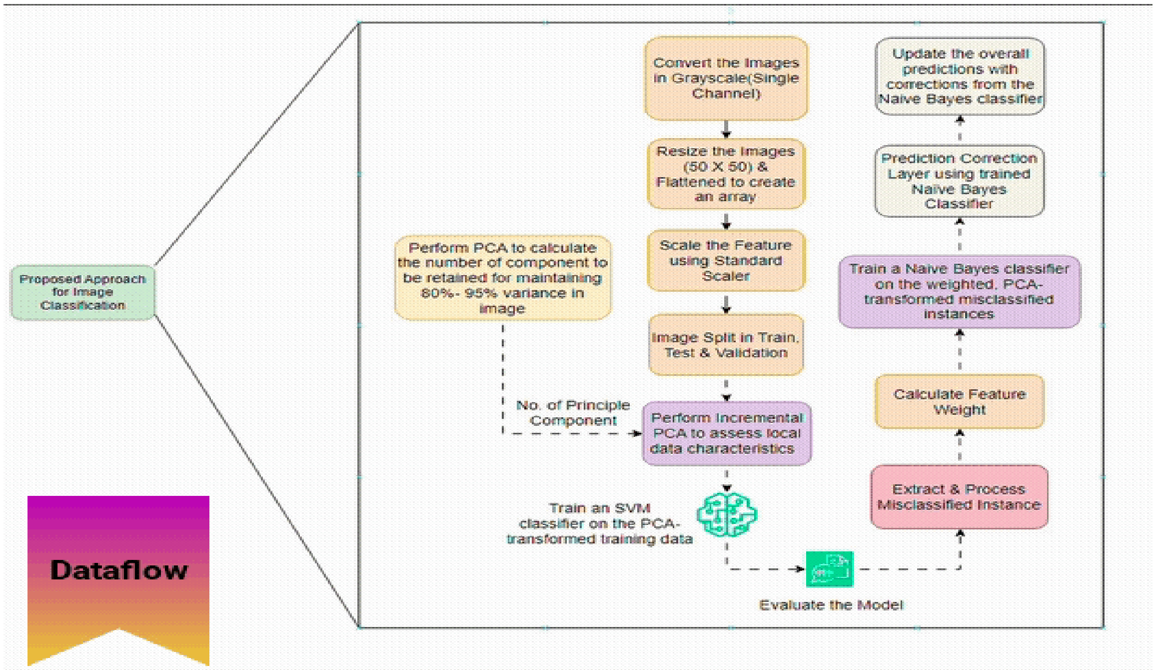


Figure 5.2: Expanded view of Proposed Architecture of Image Classification

❖ Incremental PCA for Feature Selection

The proposed method initiates with Incremental PCA, wherein the number of principal components is adaptively optimized for different subsets of the data. Unlike traditional PCA which applies a fixed number of components globally, Incremental PCA assesses local data characteristics to determine the optimal number of components. This adaptive dimensionality reduction ensures that critical features are retained, potentially enhancing the classifier's performance. By focusing on localized data structures, Incremental PCA offers a more nuanced approach to dimensionality reduction, particularly beneficial for datasets with heterogeneous characteristics.

❖ Hierarchical Ensemble Classifier

To further refine classification accuracy, we introduce a Hierarchical Ensemble Classifier comprising two levels:

- **Level 1**: An initial classification is performed using PCA followed by SVM. PCA reduces the dimensionality of the dataset, and SVM provides a robust initial classification.

- **Level 2**: Instances misclassified by the SVM are subjected to a secondary PCA, which focuses on different principal components that might capture additional variance not considered in the initial PCA. These reclassified instances are then processed using a Naive Bayes classifier. The PCA-transformed data for these misclassified instances is scaled using feature weights derived from the initial PCA, enhancing the model's focus on more significant features during reclassification.

❖ **PCA-Enhanced Naive Bayes with Feature Weights**

Rather than altering the internal mechanics of the Naive Bayes classifier, the enhancement involves scaling the PCA-transformed data of misclassified instances using feature weights. These weights are derived from the sum of absolute values of the components from the initial PCA, reflecting the importance of each feature. By scaling the misclassified instances' data with these weights before applying Naive Bayes, the approach emphasizes features deemed significant by the PCA, thereby indirectly influencing classification results. This method maintains the simplicity and computational efficiency of Naive Bayes while leveraging the insights from PCA to improve accuracy.

5.3.2 Adjustment in the proposed model

The proposed model, once trained on the available dataset, demonstrated its capability to effectively extract features and perform significantly well on data of a similar nature. The model is characterized by its compact size and high training speed, which are notable strengths. However, its performance on whether the "image is panel or not" is not improved because proposed model was trained exclusively on panel images, without exposure to other types of images. As a result, while it performs exceptionally well on similar panel images due to effective feature extraction, it struggles significantly when presented with non-panel images. This limitation fail to serve our purpose of creating single model to fulfil the dual purpose of PV panel identification & PV panel fault identification.

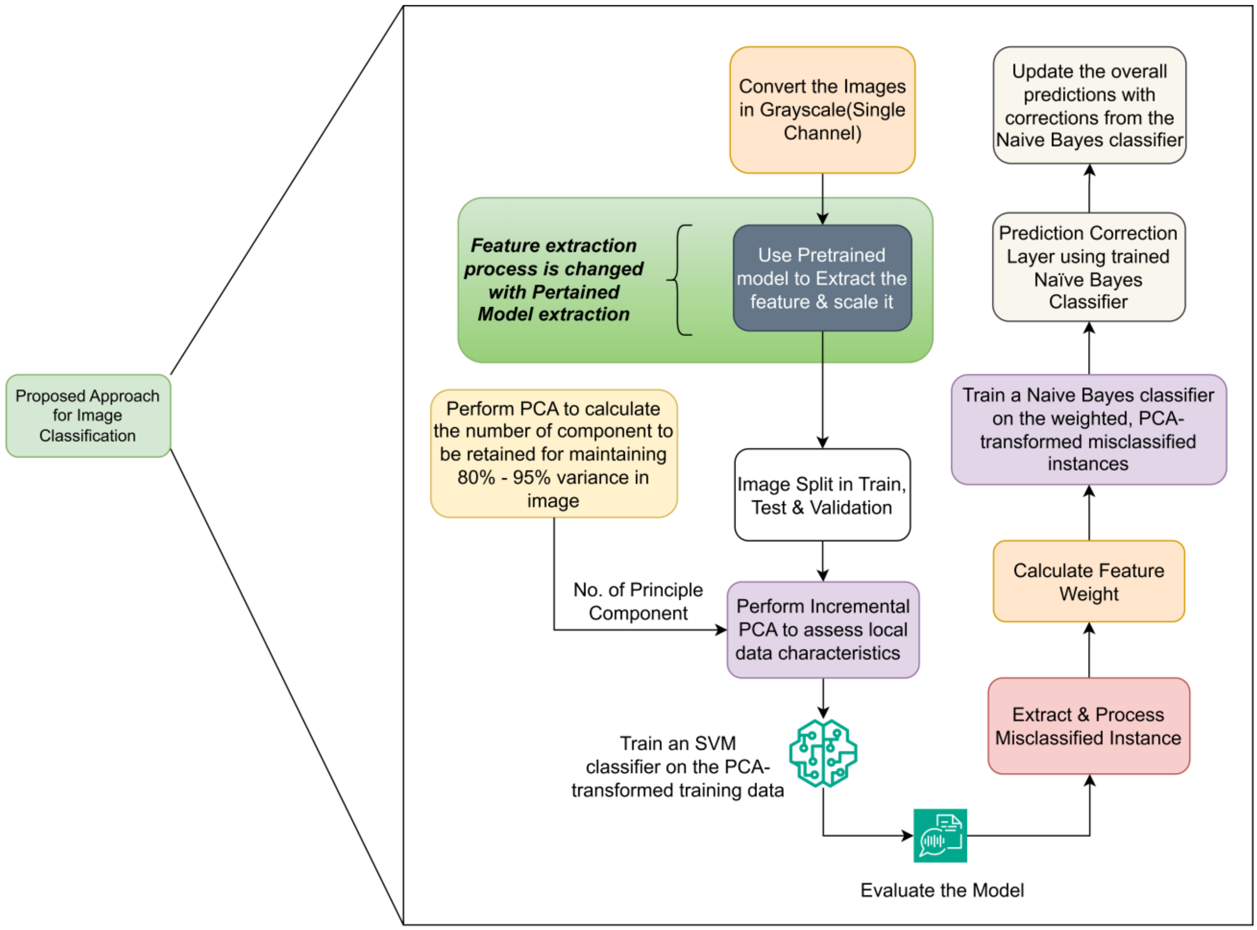
This limitation arises because the feature extraction process during training was tailored specifically to panel images, preventing the model from generalizing effectively to different types of data.

Here are the few approaches we can take to solve this problem,

- **Approach 1:** Train our model with large set of data or trained it using image net data and see whether we can generalize it or not.
- **Approach 2:** Alternatively, leverage a pre-trained model to utilize its extracted features for further analysis.

5.3.3 Modification in Proposed Model

While attempting to improve the model performance using Approach 1, we faced challenges with infrastructure availability. Since most of the model training was conducted using Google Colab's GPU/CPU, the varying resource availability made it difficult to complete the training. Consequently, we had to rely on Approach 2, leverage the pre-trained model and its extracted features for our subsequent analysis and efforts to enhance the model's performance.


Proposed Approach for Image Classification
Convert the Images in Grayscale(Single Channel)
Feature extraction process is changed with Pertained Model extraction
Use Pretrained model to Extract the feature & scale it
Perform PCA to calculate the number of component to be retained for maintaining 80% - 95% variance in image
Image Split in Train, Test & Validation
No. of Principle Component
Perform Incremental PCA to assess local data characteristics
Train an SVM classifier on the PCA-transformed training data
Evaluate the Model
Extract & Process Misclassified Instance
Calculate Feature Weight
Train a Naive Bayes classifier on the weighted, PCA-transformed misclassified instances
Prediction Correction Layer using trained Naïve Bayes Classifier
Update the overall predictions with corrections from the Naive Bayes classifier

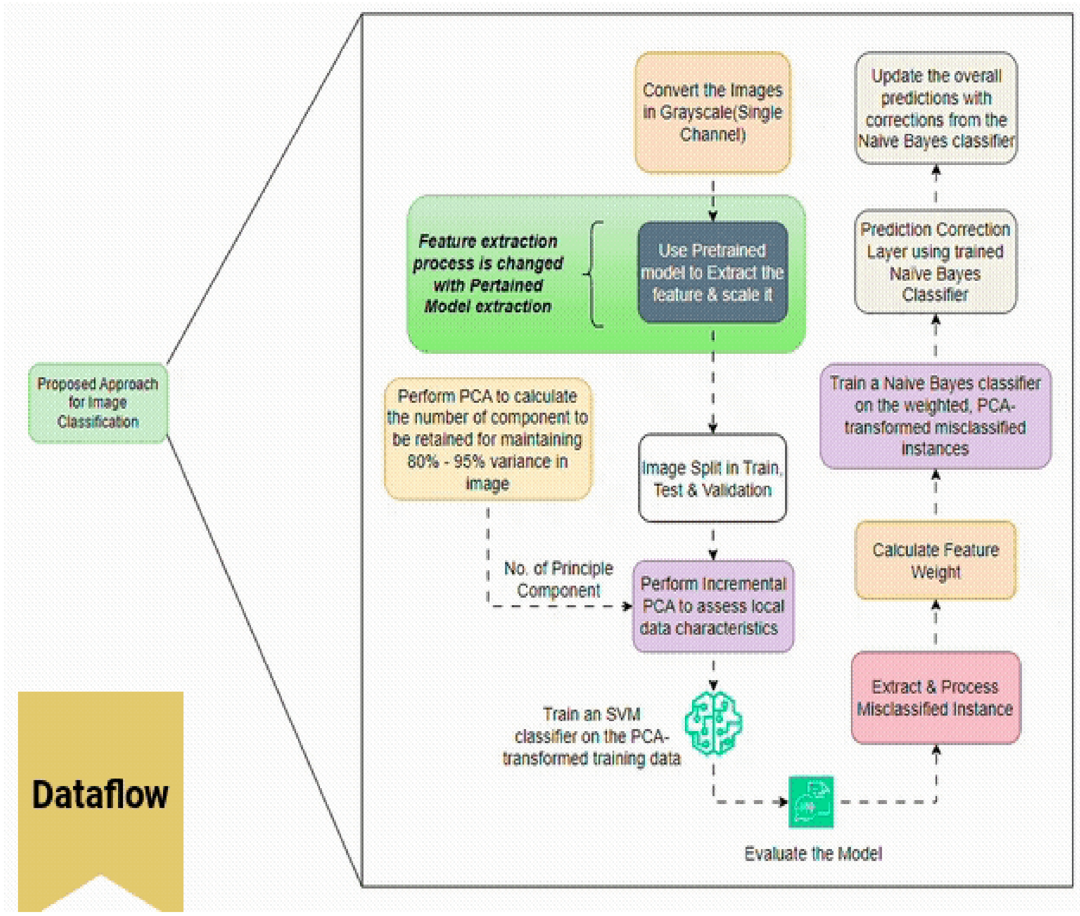


Figure 5.3: Adjusted Proposed Architecture of Image Classification

Fig 5.3 shows the adjustment in the proposed model highlighted in shaded green box where the feature selection is now happening with pre-trained model.

This adjusted architecture works very well, and accuracy is also maintained for both seen as well as unseen data. This also serve our purpose of creating single model to fulfil the dual purpose of PV panel identification & PV panel fault identification.

5.3.4 Rationale behind the improvement in the Adjusted Proposed Architecture

The proposed methodology for image classification of photovoltaic (PV) panel faults employs a novel integration of classical machine learning techniques, namely Incremental Principal Component Analysis (PCA), Support Vector Machine (SVM), and Naive Bayes, supplemented by the strategic use of a pretrained model for initial feature extraction. This hybrid approach capitalizes on the strengths of each technique, optimizing the feature extraction process by leveraging the advanced capabilities of deep learning models pretrained on extensive datasets, thus ensuring a robust initial set of features. Further refinement is achieved through PCA, which reduces the dimensionality of the data while retaining 95% of the variance, thereby enhancing computational efficiency and focusing the SVM classifier on the most salient features for fault detection. The integration of an incremental PCA ensures that the system dynamically adapts to new data, maintaining high classification accuracy in diverse operational environments.

The rationale behind the exemplary performance of this classification system is anchored in its advanced error correction mechanism, where a Naive Bayes classifier is specifically trained on instances initially misclassified by the SVM. This targeted approach allows for the continual refinement of the model by learning from its mistakes, significantly reducing false positives and negatives over time. Additionally, the system incorporates a feature weighting and correction layer, which recalibrates the decision boundaries based on the insights gained from misclassified instances, thereby enhancing the overall predictive accuracy. This methodology not only increases the reliability of fault detection in PV panels but also offers a scalable and adaptive solution that can be extended to other complex image classification tasks, highlighting its utility and innovative application in real-world scenarios.

5.4 PV Fault Classification using Tabular Fault Data

5.4.1 Tabular Fault analysis using Logistic regression and SGD classifier fusion

As a continuation of our fault classification task, we are now focusing on classifying faults in a simulated 250-kW PV power plant. The data for this task includes voltage and current parameters and consists of three types of faults.

After an initial analysis and preliminary checks on the data, we decided to approach this problem as a multiclass classification task using a purely machine learning-based method.

5.4.1.1 Initial Approach: Logistic Regression

Based on historical success of Logistic Regression in classification task we first implemented Logistic Regression in our data, which yielded good classification accuracy on the available dataset. However, to meet our future goals of exploring incremental learning for fault classification, we also experimented with Stochastic Gradient Descent (SGD) along with Logistic Regression.

5.4.1.2 Incremental Learning with SGD

Though the result of both this approach looks equivalent on the dataset, SGD is particularly useful for optimizing the Logistic Regression model iteratively, allowing it to adapt quickly to new data. The benefits of using SGD in line with incremental training and improved classification include:

1. **Incremental Training:**
    - The SGD Classifier updates the Logistic Regression model iteratively.
    - This ensures that the model stays current with the latest data, maintaining high performance in dynamic operational environments.
2. **Performance Monitoring:**
    - We continuously monitor the model's performance on a validation set to detect any signs of overfitting or underfitting.
    - Necessary adjustments are made to maintain the model's effectiveness.

Incremental learning is described in section 5.6 under chapter 5

5.4.1.3 Steps Followed in Our Approach without streaming data

1. **Data Analysis:** Conduct initial analysis and checks on the tabular data from the PV power plant.
2. **Model Selection:** Choose Logistic Regression for initial classification.
3. **Performance Monitoring:** Montor the performance on the data and report classification result in confusion matrix.

Steps followed in this approach is depicted below,

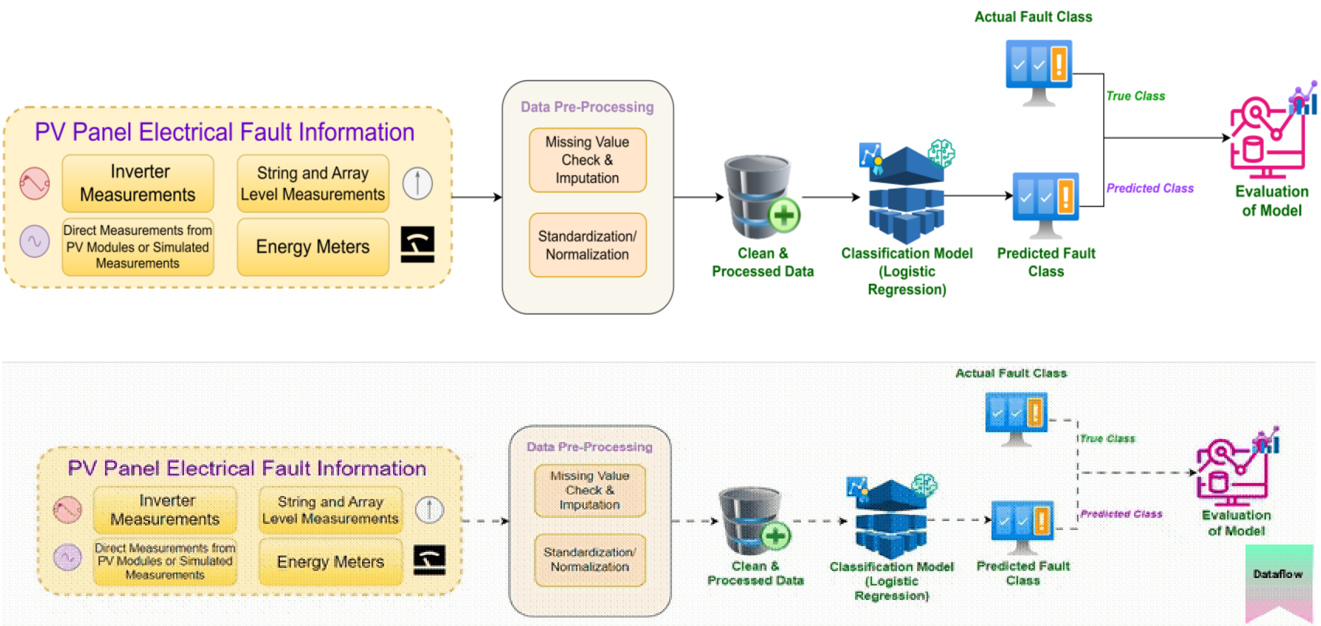


Figure 5.4: PV Panel Tabular Fault Classification

5.5 Multimodal Approach

5.5.1 Overview of Decision Level Fusion

Late Fusion also known as Decision Level Fusion is an approach involves running both models independently on their respective datasets and then combining their predictions at the decision level. Here are a few ways to implement late fusion:

- **Voting Scheme**: Each model's output can be treated as a vote. The final decision is made based on majority voting if there are multiple models or a simple logical rule if there are just two models.
- **Score Averaging**: If both models produce probabilistic outputs or scores, you can average these scores. Optionally, you can weight these averages if one model is known to be more reliable.
- **Rule-Based Decision**: Develop specific rules that consider the outputs from both models. For example, if the image model detects a defect and the tabular model indicates a high likelihood of failure, the final decision could be to classify the observation as high risk.

5.5.2 Our Scenario & Approach

We are addressing the challenge of classifying faults in photovoltaic panels using two different types of data, images and tabular data. To effectively handle this, we have developed a decision-level fusion multimodal approach. This approach will allow us to classify faults by leveraging both image data and tabular data.

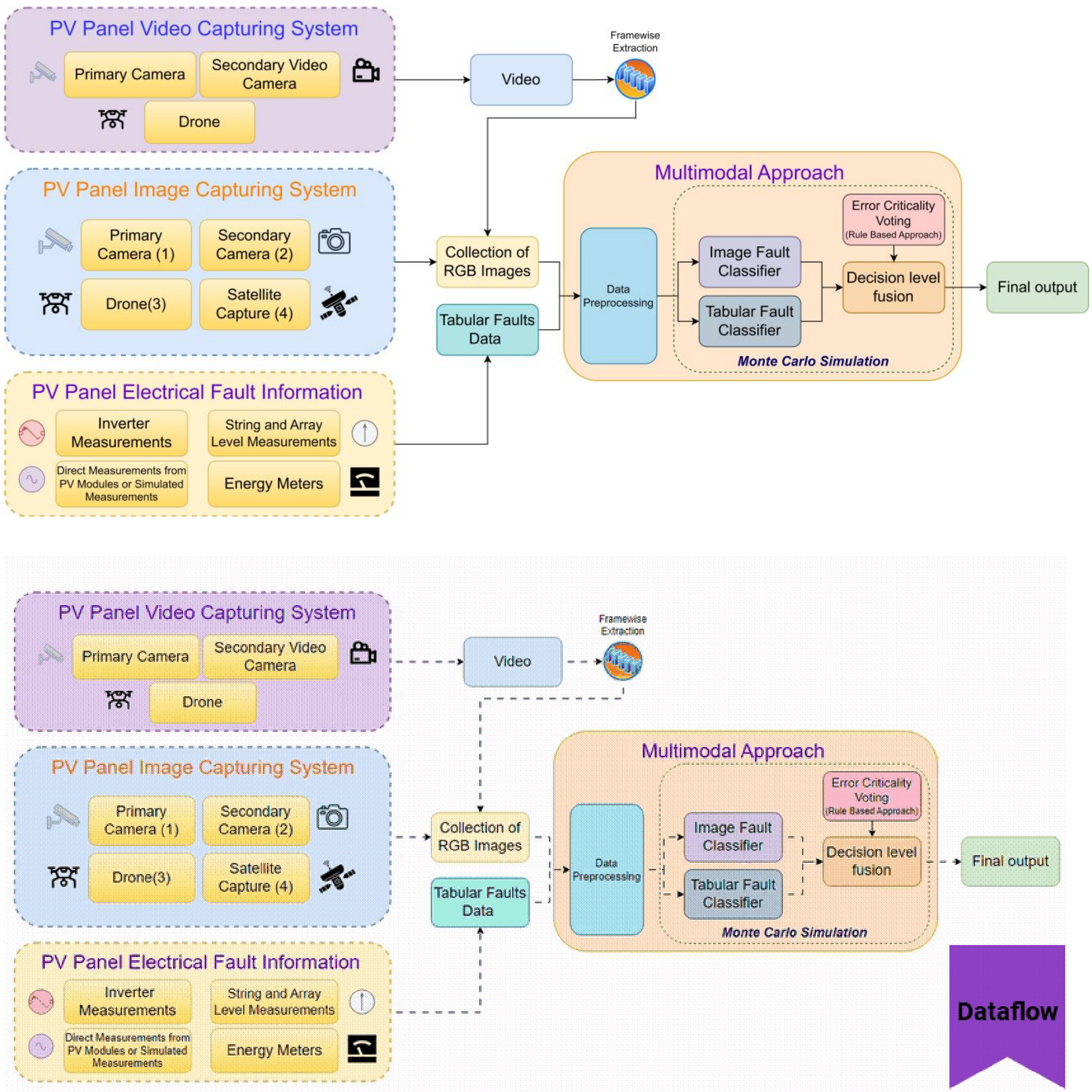


Figure 5.5: Multimodal Approach in Fault Classification

Approach Overview

1. **Independent Models for Each Data Type**
    - **Image Model:** We have a model specifically designed to classify faults based on image data. This model analyzes visual features and identifies issues visible in the images.
    - **Tabular Model:** We also have a separate model that focuses on classifying faults using tabular data. This model processes structured data such as numerical values and categorical information related to the photovoltaic panels.
2. **Decision-Level Fusion**
    - Since we are using two different models for two different types of data, we need a mechanism to combine their outputs effectively. This is where the decision-level fusion comes in.

Integrating the Models with a Voting Scheme

In scenarios where both image data and tabular data are available, we can use a voting scheme to integrate the outputs of the two models. Here's how this can be structured:

1. **Separate Classification:** The image model and the tabular model process their respective inputs and produce individual classification results.
2. **Decision Combination:** We then combine these results using a voting scheme using error criticality voting and Monte Carlo simulation result of prediction. This can be implemented as below:
    - **Error Criticality Voting:** Each fault in both image and tabular faults has some criticality level assigned to it refer Table 5.1, and based on the output which ever fault has high criticality, get majority and corresponding results are displayed.
    - **Monte Carlo Simulation:** To enhance this process, we incorporate Monte Carlo simulations to estimate the probabilities of different fault types and improve the robustness of our final decision. By running multiple

simulations in inference level, we can better account for variabilities in the data, leading to more reliable classification results.

Detailed Workflow

1. **Data Input:** The system receives data which could be either images, tabular data, or both.
2. **Model Processing:**
    - **Image Data:** If image data is provided, it is processed by the image model to classify faults.
    - **Tabular Data:** If tabular data is provided, it is processed by the tabular model to classify faults.
    - **Both Image and Tabular Data:** Model identify the data type and take a decision on selecting the necessary model to classify the fault.
3. **Decision Fusion:** If both types of data are available:
    - The image model and the tabular model each produce a classification result.
    - The results are combined using an error criticality rule and with Monte Carlo simulation at inference, produce the final classification.
4. **Final Output:** The system outputs the final classification based on the fused decisions from both models explained in decision combination section.

This multimodal approach ensures that we can effectively classify faults in photovoltaic panels by leveraging the strengths of both image data and tabular data. By combining the outputs of the two specialized models, we achieve a more robust and accurate fault classification system.

Error Criticality Voting a Rule based approach:

- **Image Faults**: Bird Drop, Dusty, Snow Cover, Electrical Faults, Physical Damage

- **Tabular Faults**: Class 0, Class 1, Class 2, Class 3

Table 5.1: Criticality & Priority for Faults

Fault Type	Criticality Level	Priority Level (1-5)
Physical Damage	High	5
Electrical Faults	High	5
Snow Cover	Medium	3
Dusty	Medium	3
Bird Drop	Low	1
Class 3	High	5
Class 2	Medium	3
Class 1	Low	2

Image Faults   Tabular Electrical Faults

5.6 Incremental Learning, Feedback Mechanism & Automated Retraining

5.6.1 Overview of Incremental Learning

Incremental learning is a crucial method for continually enhancing the performance of fault classification models in solar power systems. As the adoption and importance of renewable energy, particularly solar power, grow, ensuring the optimal operation of solar power plants becomes essential. Faults in photovoltaic (PV) panels can significantly impede the efficiency of these power plants, making rapid and accurate fault detection and classification vital for maintaining high energy output and system health.

5.6.2 Role of Feedback Mechanism & Incremental Learning

The feedback mechanism serves as a dynamic tool to integrate real-world insights into the machine learning models responsible for detecting and classifying faults in PV systems. This process allows the models to adapt and improve over time, reflecting changes in environmental conditions, new types of faults, or variations in panel behaviour.

5.6.3 How Feedback Mechanism Works

After data collection and pre-processing the data is passed through the desired model based on its behaviour and once the initial training and evaluation is completed, initial set of models is ready for use. But to handle variation of incoming data and to maintain the accuracy of the model, we have to rely on the user feedback mechanism.

The base network and incrementally improved network details are shown in Fig 5.6

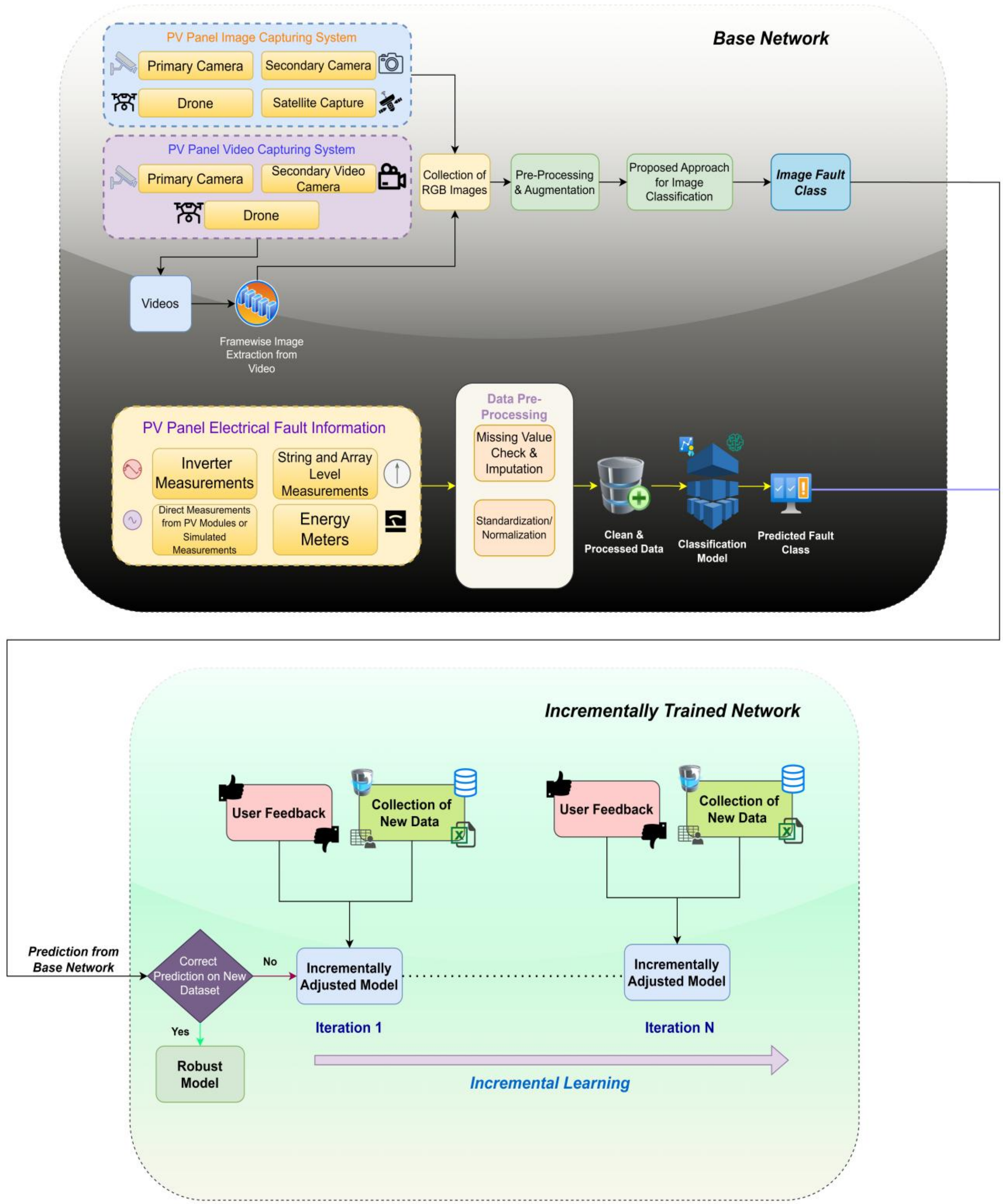


Figure 5.6: Base Network & Incrementally Trained Network Architecture

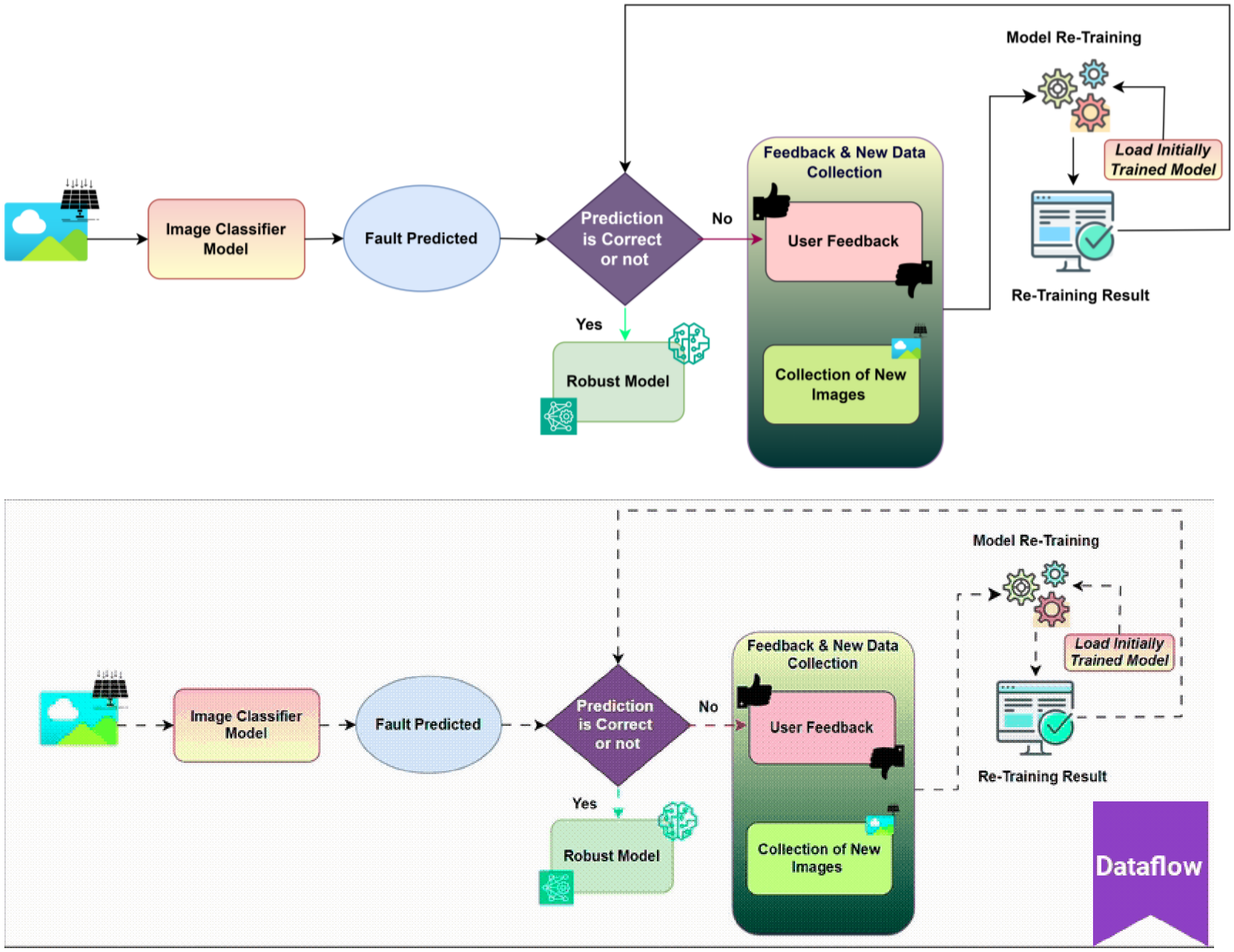
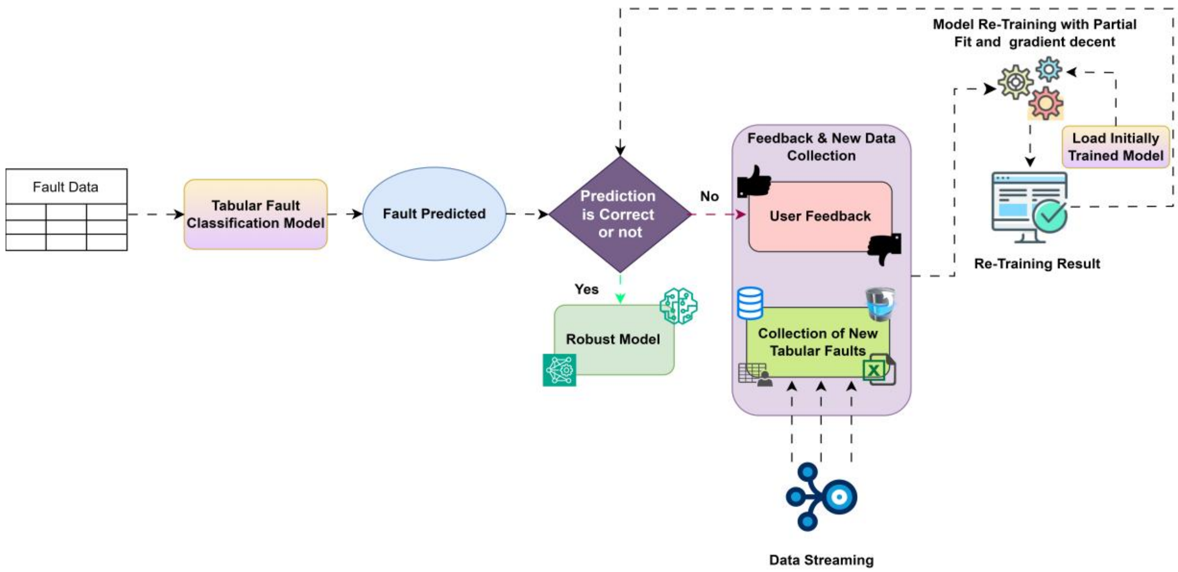


Figure 5.7: Image Data: Incrementally Trained Network Architecture



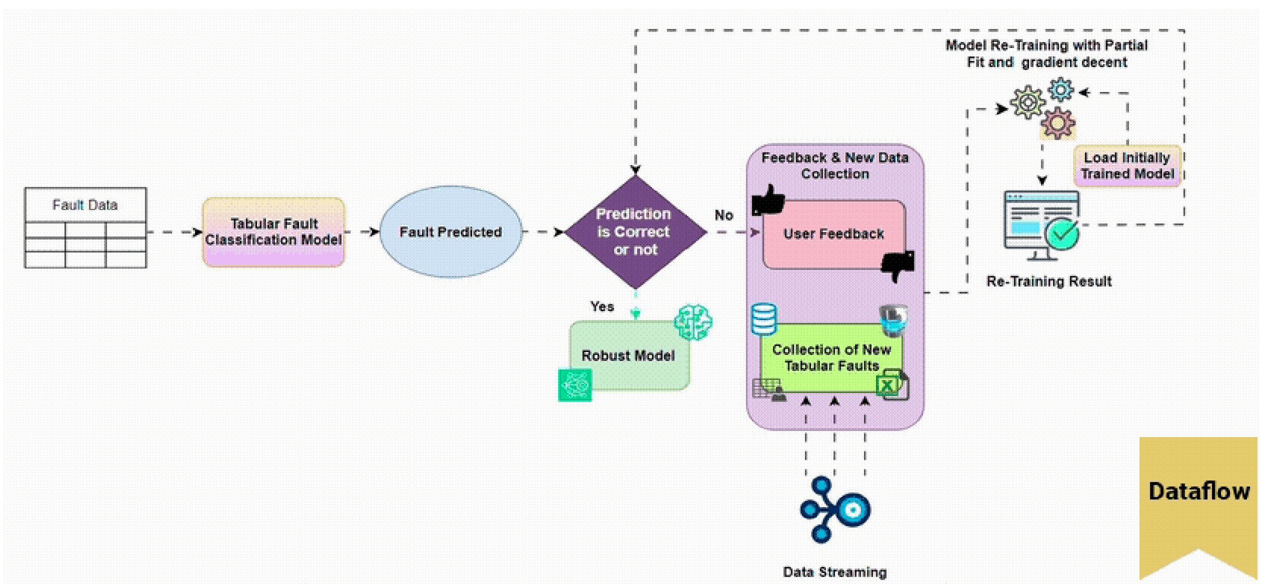


Figure 5.8: Tabular Fault Data: Incrementally Trained Network Architecture

5.6.3.1 Integration of User Feedback:

- **Feedback Capture:**
  - Users (maintenance personnel or system operators) flag misclassifications through a dedicated interface in the system's monitoring software.
  - Feedback includes corrections to the fault classification and, for images, may also include updated annotations.

- **Feedback Processing:**
    - For image data, incorrect classifications are reviewed, and the model is retrained on these instances with corrected labels.
    - For tabular data, misclassifications are analysed to identify trends or systematic errors, and the model parameters are adjusted accordingly.

5.6.3.2 Model update process:

- **Iterative Retraining:**
    - Models are updated iteratively with new batches of feedback-enhanced data.
    - Retraining cycles are scheduled based on the accumulation of a significant number of feedback instances to balance between model stability and adaptiveness.
- **Performance Evaluation:**
    - After each update, the model's performance is evaluated on a separate validation set to monitor improvements in accuracy and reduction in misclassifications.
    - Adjustments are made to the learning rate and other training parameters based on performance metrics.

5.6.3.3 Benefits of the Feedback Mechanism & Incremental learning:

- **Enhanced Model Accuracy:** Continuous feedback integration helps refine the fault classification models, leading to higher accuracy and reliability in fault detection.
- **Adaptiveness to New Conditions:** As new fault types emerge or as operational conditions change, the feedback mechanism allows the models to adapt, ensuring they remain effective under varied conditions.
- **Reduced Downtime and Maintenance Costs:** More accurate fault detection minimizes system downtime and reduces unnecessary maintenance, thereby saving costs and enhancing the overall efficiency of solar power plants.

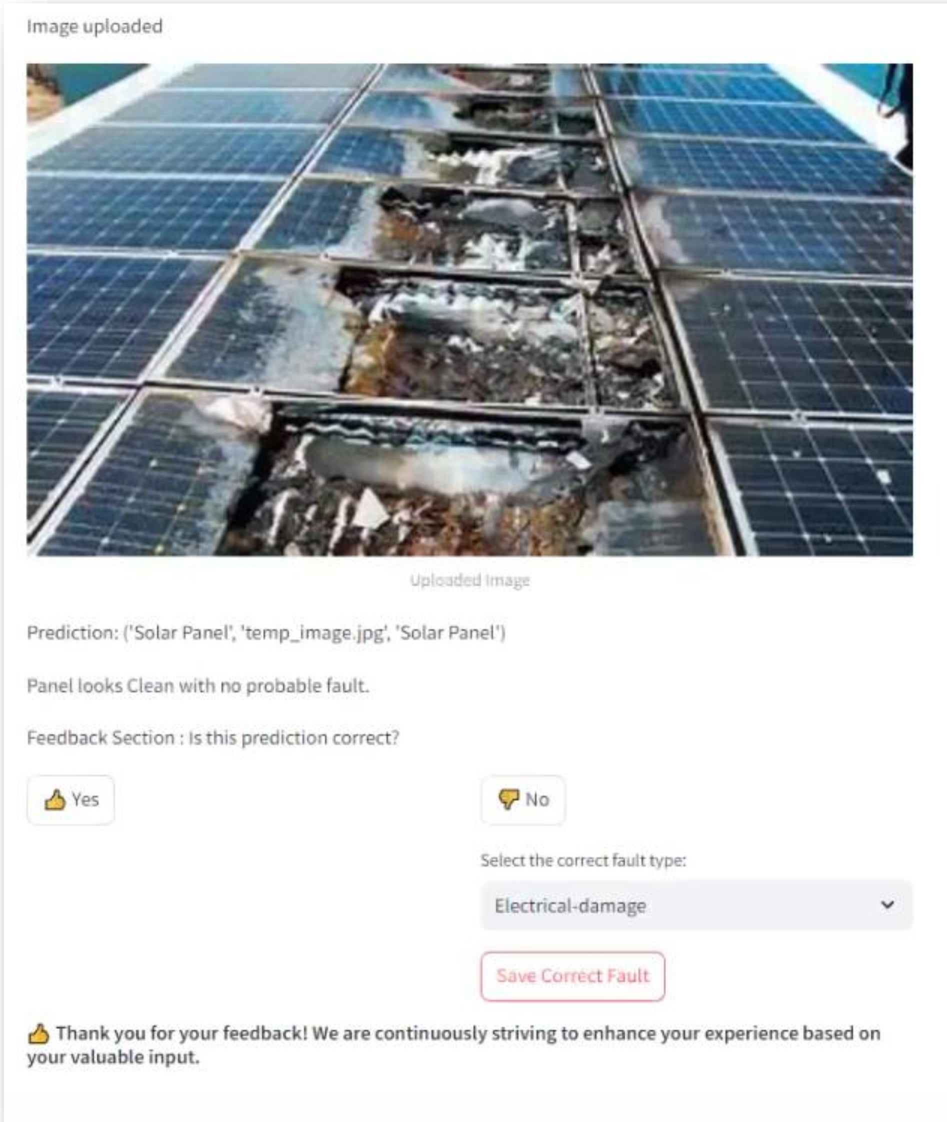


Figure 5.9: User Interface for Feedback Mechanism

5.7 Further Planning**

Below are the plan for coming weeks till Mid of August considering 15th August is the submission date,

- PV Panel fault segmentation and percentage of damage area.

*** This plan is subject to change based on future developments and circumstances.*

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 Overview

This chapter provides a comprehensive examination of the experimental evaluation conducted to assess the predictive accuracy of various classification models for PV panel identification and fault classification. The primary focus is on making a robust model which is computationally low but have potential to achieve higher accuracy if trained with large amount of data.

6.2 Environment & Settings

Proposed models build and its corresponding tuning were done using Python 3.10 and Tensorflow library in Google Colab. This platform provides GPU for a limited time, which needs to be enabled by changing the runtime for the Hardware accelerator to 'T4 GPU', which comes with 15GB memory and 12.7 GB system RAM for the free user.

6.3 Approach Taken for PV Panel Identification & Fault Classification

Primary focus of this experiment is to identify whether the given image is a photovoltaic(PV) panel image or not then find out the fault of PV panel and finally give notification to the plant operator if fault gets detected. For achieving this goal, we have taken approach of collecting diverse sets of images and videos of PV panel consisting of different types of faults and then perform some analysis on top of it and classify the faults.

We have taken reverse approach before finalizing the model architecture.

6.3.1 Exploring Different CNN Based Architecture

Based on the literature study we have noticed that different deep learning-based CNN architecture perform very well on classifying the imaged based faults. Relying on this approach or convention we also start exploring different CNN based architecture.

We have explored below CNN based architecture on our PV panel faulty image dataset,

Table 6.1: Explored CNN based architecture

CNN based Model Names		
▪ Xception ▪ NASNetLarge ▪ Vgg-16	▪ ResNet50 ▪ MobileNetV2 ▪ InceptionV3	▪ MobileNetV3 ▪ NASNetMobile ▪ DenseNet

On top of trying above models we also perform hyperparameter tuning to observe the variation of the classification performance keeping in mind of creating single model for dual purpose of panel identification and further panel fault classification.

6.3.1.1 CNN Hyperparameter Tuning & Exploration

Based on the image dataset and random sampling of those images for training these CNN based architecture leads to varying accuracy over different sampling. So, to find out which sample of images leads to good accuracy after training on test or unseen data is a challenge. To, analyses it we have done couple of modification on batch size and random seed to change the sampling image size and image set and train the same CNN based architecture repeatedly to see the variation in the accuracy and noted it.

6.3.1.1.1 Tuning parameter combination

- **Batch Size:** 4, 8, 16, 32, 64, 120
- **Epochs:** 5, 10, 15, 20
- **Random Seed:** 30, 40, 42, 45, 50, 55, 56, 69, 120
- **Optimizer**: Adam, SGD, AdaGrad, RMSProp
- **Learning Rates**: 0.1,0.01,0.001,0.0001
- **Dropouts**: 0.1, 0.3, 0.5, 0.7, 0.8, 0.9

6.3.1.1.2 Varying Main Model on top of Base CNN Model

Table 6.2 explains the two different architectures used on top of base CNN models to be performing classification.

Table 6.2: Main Model on top of CNN Models

Model 1	Model 2
The first model is more straightforward, leveraging the CNN model features with minimal additional layers, focusing on global pooling and dropout for regularization. It is suitable for tasks where the pre-trained features are sufficient, and minimal fine-tuning is needed.	The second model offers greater flexibility and fine-tuning capability by unfreezing the CNN model layers and incorporating a more complex sequential structure with a higher dropout rate. This model is apt for scenarios requiring fine-tuning of the pre-trained network to the specific dataset.

Conducting comprehensive hyperparameter tuning for numerous CNN models, particularly high-end architectures such as InceptionV3 and NASNetLarge, is an inherently time-consuming process. However, through systematic evaluation of various hyperparameter combinations, we have successfully identified the optimal set of parameters that yield the best performance for each model on the given dataset. The table below provides a summary of these best-performing configurations among which Xception performed best having accuracy over 99% on training data and over 95% on the test data is recorded.

Table 6.3: Best Result using CNN Models for Panel Fault Classification

Panel Fault Classification									
Base Model	***Main Model***	***Seed***	***Batch Size***	***Epochs***	***Optimizer***	***Learning Rates***	***Training Accuracy***	***Validation Accuracy***	***Test Accuracy***
Vgg-16	Model 2	56	32	10	Adam	0.001	0.9859	0.9541	0.9456
DenseNet	Model 2	30	64	10	Adam	0.001	0.9743	0.9356	0.9054
InceptionV3	Model 2	55	64	15	Adam	0.001	0.9751	0.9458	0.9173
Xception	***Model 2***	***42***	***32***	***15***	***Adam***	***0.0001***	0.9975	0.9552	0.9548
ResNet50	Model 2	30	32	10	Adam	0.01	0.9847	0.9695	0.9281
MobileNetV2	Model 2	42	32	15	Adam	0.001	0.9284	0.8974	0.8546
MobileNetV3	Model 2	42	32	10	Adam	0.001	0.9548	0.9398	0.8594

NASNetMobile	Model 2	30	32	10	Adam	0.001	0.9854	0.9547	0.9005
NASNetLarge	Model 2	30	32	5	Adam	0.01	0.9953	0.9721	0.9473

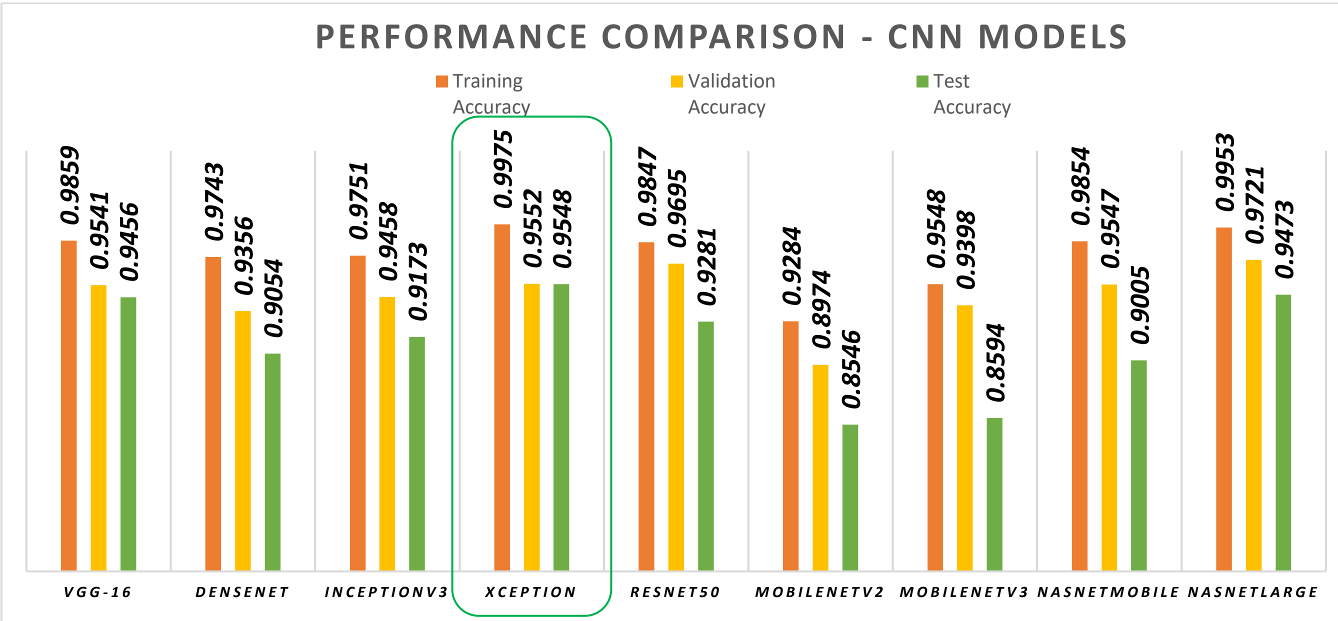


Figure 6.1: Performance Comparison of different CNN Models: Panel Fault Classification

6.3.1.1.3 Challenges

Below are the major challenges which leads us to search for other alternative approaches,

- Training time is very high and computationally very heavy.
- CNNs have many hyperparameters (e.g., learning rate, batch size, number of layers, dropout rate) that significantly impact performance. Finding the optimal set of hyperparameters is often a time-consuming and complex task.
- CNN is prone to overfitting when trained with small set of data.

6.3.2 Exploring Machine Learning Based Approach

Initial attempt to use CNN based architecture works well in classification but to improve the training time and to reduce model complexity we have explored pure machine learning based support vector machine algorithm.

6.3.2.1 SVM Hyperparameter Tuning & Exploration

The results presented in Table 6.3, obtained using the Support Vector Machine (SVM) algorithm, represent the optimal outcomes achieved following an extensive hyperparameter tuning process. The specific details of this tuning are delineated in Table 6.4.

Table 6.4: Hyperparameter for SVM

Hyperparameter	***Selected Value***	***Explored Values***	***Explanation***
Kernal	poly	rbf, poly, linear	Specifies the kernel type to be used in the algorithm, 'rbf' stands for Radial Basis Function, suitable for non-linear problems.
Gamma	Scale		Kernel coefficient for 'rbf', 'poly', and 'sigmoid', where 'scale' uses 1 / (n_features * X.var()) as value by default.
C	0.5	0.1,0.2,0.5, 0.8,0.9,1.0	Regularization parameter controlling trade-off between achieving a low error on training data and minimizing the norm of the weights.

Experimental results of hyperparameter tuning is shown in in below Table 6.5

Table 6.5: Hyperparameter Tuning - SVM Model Results for Panel Fault Classification

Panel Fault Classification							
Model	***Seed***	***Kernal***	***Gamma***	***C***	***Training Accuracy***	***Validation Accuracy***	***Test Accuracy***
SVM	42	Poly	Scale	0.5	0.798	0.74506	0.7447
SVM	42	rbf	Scale	0.5	0.905	0.5986	0.5743
SVM	42	Linear	Scale	0.5	0.7289	0.6541	0.6258
SVM	42	Poly	Scale	0.1	0.755	0.715	0.705

SVM	42	Poly	Scale	0.2	0.77	0.7246	0.7153
SVM	42	Poly	Scale	0.8	0.8231	0.764	0.7225
SVM	42	Poly	Scale	0.9	0.7985	0.7211	0.6985
SVM	42	rbf	Scale	0.1	0.85	0.6589	0.6235
SVM	42	rbf	Scale	0.2	0.875	0.6347	0.6147
SVM	42	rbf	Scale	0.8	0.92	0.5874	0.5547
SVM	42	rbf	Scale	0.9	0.9365	0.5748	0.5478
SVM	42	rbf	Scale	1	0.9487	0.5626	0.5369
SVM	42	Linear	Scale	0.1	0.7158	0.6687	0.6198
SVM	42	Linear	Scale	0.2	0.715	0.658	0.6489
SVM	42	Linear	Scale	0.8	0.7482	0.652	0.6124
SVM	42	Linear	Scale	0.9	0.7451	0.65	0.6149
SVM	42	Linear	Scale	1	0.7563	0.6415	0.5895

This algorithm significantly mitigates the issue of high training time, achieving a reduction of approximately 85% due to its simpler architecture. However, this simplification comes at the cost of performance. Specifically, we observe a decline in training accuracy by nearly 20%, and a decrease in validation and test accuracies by approximately 22% when compared to the best-performing CNN model in our study, which is the Xception model.

Table 6.6: Best Result using SVM Model for Panel Fault Classification

Panel Fault Classification							
Model	***Seed***	***Kernal***	***Gamma***	***C***	***Training Accuracy***	***Validation Accuracy***	***Test Accuracy***
SVM	42	Poly	Scale	0.5	0.798	0.74506	0.7447

6.3.3 Exploring Proposed Approach

While utilizing Support Vector Machines (SVM), we observed a significant acceleration in the training process. Although this approach effectively addresses our computational efficiency concerns, it comes at the expense of performance. Consequently, this necessitates further exploration of alternative methods that can simultaneously achieve computational efficiency and high performance.

Leveraging the power of machine learning, we have developed the proposed approach outlined in Figures 5.2 and 5.3, which demonstrates not only computational efficiency but also enhanced performance. By utilizing pre-trained models for feature extraction and subsequently integrating various independent machine learning techniques, as detailed in Chapter 5, we have achieved high accuracy comparable with CNN models but with less computational efficiency.

6.3.3.1 Exploration Criterion alongside Proposed Approach

Exploration and test results are compared by varying below parameters in different stage of our proposed approach,

- Percentage of Variance (0.80 - 0.95)
- SVM Hyperparameter explained in Table 6.4
- Feature extraction using 10 pretrained model
- Explore transformer-based feature extractor

Initially via thorough exploration, we determined that capturing 80% of the variance is sufficient to achieve satisfactory accuracy in fault classification. Consequently, we have reduced the total number of principal components to 117.

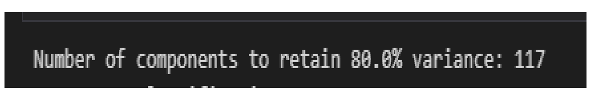


Figure 6.2: Principle component analysis results

We selected a polynomial kernel for the Support Vector Machine (SVM) algorithm, with a regularization parameter CCC set to 0.5 and a gamma value configured as 'scale'. This parameter configuration was determined to be the most effective for our dataset based on empirical evaluations and performance metrics.

6.3.3.2 Proposed Approach Results in Fault Classification

After implementing the series of steps outlined in the adjusted proposed approach and experimenting with various hyperparameters and feature extractor models, we observed significant improvements in the overall performance metrics. Fig 6.3 shows the result of our proposed approach using pretrained feature extractor model from hugging face library.

PCA + SVM Classification Report:

	precision	recall	f1-score	support
0	0.83	0.85	0.84	157
1	0.64	0.47	0.55	19
2	0.89	0.84	0.86	74
3	0.74	0.77	0.76	102
4	0.25	0.27	0.26	15
5	1.00	0.98	0.99	51
accuracy			0.81	491
macro avg	0.74	0.72	0.73	491
weighted avg	0.81	0.81	0.81	491

Final Classification Report after implementing Proposed Approach

	precision	recall	f1-score	support
0	1.00	0.98	0.99	157
1	0.86	0.95	0.90	19
2	0.99	0.99	0.99	74
3	0.97	0.95	0.96	102
4	0.72	0.87	0.79	15
5	1.00	1.00	1.00	51
...				
accuracy			0.98	491
macro avg	0.93	0.96	0.95	491
weighted avg	0.98	0.98	0.98	491

Figure 6.3: Adjusted proposed approach results on Test Dataset

Fig 6.3 explained that the proposed approach demonstrates a marked improvement in classification performance compared to the PCA + SVM method. Notably:

- **Accuracy** increased from 0.81 to 0.93.
- **Macro average precision** increased from 0.74 to 0.93.
- **Macro average recall** increased from 0.72 to 0.96.
- **Macro average F1-score** increased from 0.73 to 0.95.

The improvements are particularly significant for the previously underperforming classes, such as class 4, where the F1-score increased from 0.26 to 0.79, and class 1, where it improved from 0.55 to 0.90. These results indicate that the proposed approach not only enhances the overall accuracy but also addresses class imbalance issues and improves the model's ability to generalize across all categories.

The enhanced performance metrics across all classes suggest that the proposed approach effectively captures and leverages more discriminative features, leading to more accurate and robust classifications. This highlights the efficacy of the proposed model modifications and their potential applicability in similar classification tasks.

Below are several examples runs on the test dataset, illustrating the performance of proposed models.

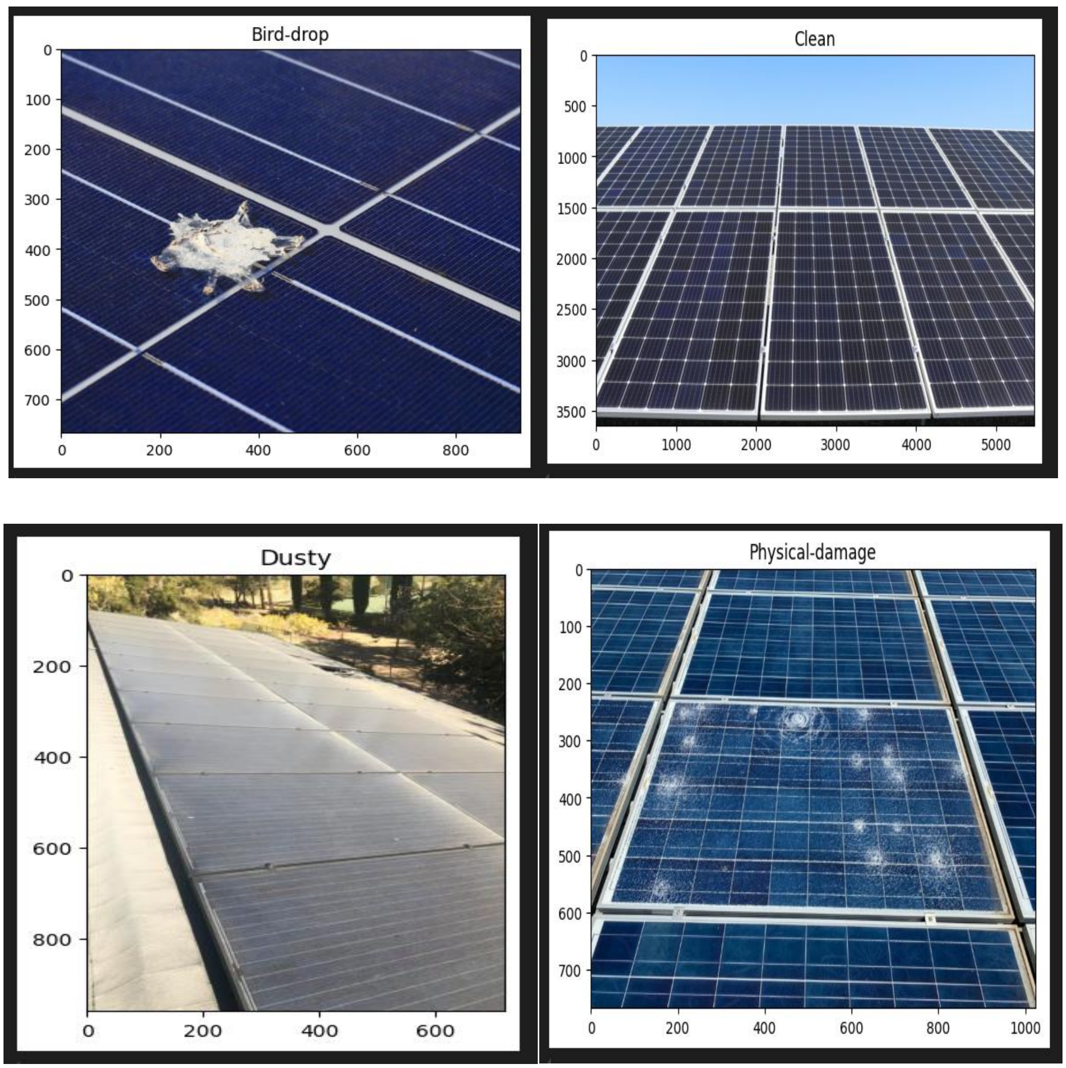


Figure 6.4: Adjusted proposed approach fault detection screenshot

6.3.4 Solar Panel Identification

In our proposed system, we aim to develop a robust method capable of identifying solar panels and subsequently detecting faults within them. Our result discussion has primarily concentrated on elucidating the efficacy of our proposed approach in identifying faults in photovoltaic (PV) panels and the methodology through which this approach was formulated.

To further assess the capability of the proposed approach, we trained the model using approximately 590 images of non-PV panels sourced from the Pexel website, alongside available PV panel images. This experiment was conducted to evaluate the accuracy of the classification task and validate the robustness of our proposed method in panel identification.

- Train & Validation set : 325 images
- Test set : 265 images

Proposed model works extremely well in classifying the PV solar panel as shown in Fig 6.5 and this conclude that we can utilize the same novel approach for serving dual purpose of solar panel identification and further to that solar panel fault detection.

```
PCA + SVM Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.77      0.83        56
           1       0.94      0.98      0.96       209

    accuracy                           0.94       265
   macro avg       0.93      0.87      0.90       265
weighted avg       0.93      0.94      0.93       265

Final Classification Report after implementing Proposed Approach
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        56
           1       1.00      1.00      1.00       209

    accuracy                           1.00       265
   macro avg       1.00      0.99      0.99       265
weighted avg       1.00      1.00      1.00       265
```

Figure 6.5: Adjusted proposed result in PV Panel classification

6.3.5 Comparative study

We have explored multiple CNN based architecture, pure machine learning based model like PCA+SVM and our proposed and have seen our proposed approach perform comparatively well on available dataset in identifying the panel and classifying there faults. Below Table 6.7 explain the comparative result of different models,

Table 6.7: Comparative Study : Performance of different models in Image Classification

Panel Fault Classification				
Model	***Training Accuracy***	***Validation Accuracy***	***Test Accuracy***	***Avg. Training Timing in Min***
CNN based Models	0.9859	0.9541	0.9456	90
ML - PCA+SVM	0.798	0.74506	0.7447	10
Proposed Approach	0.999	0.998	0.98	15

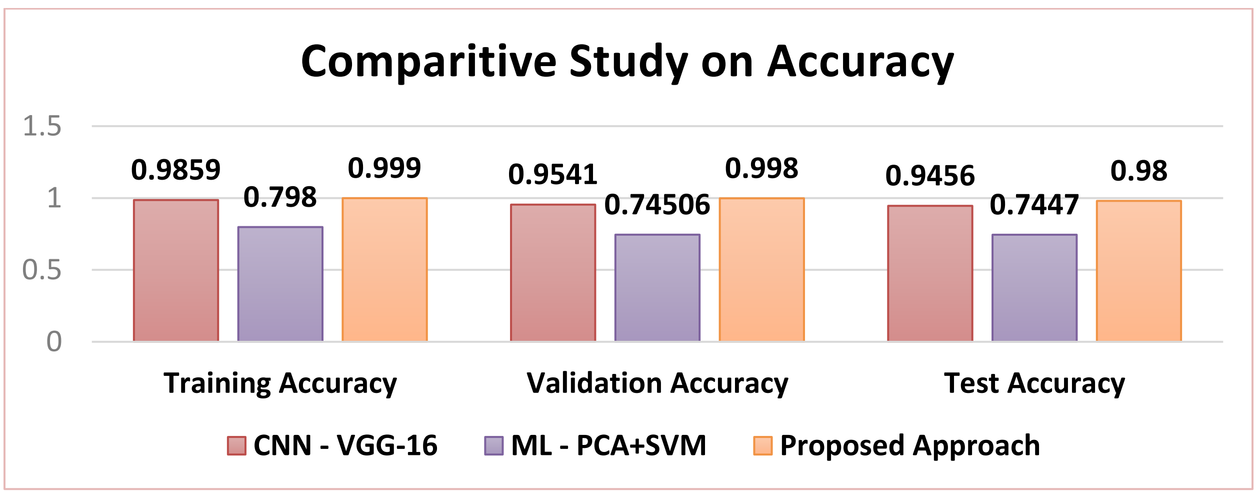


Figure 6.6: Comparative Study in Accuracy of Panel Fault Detection

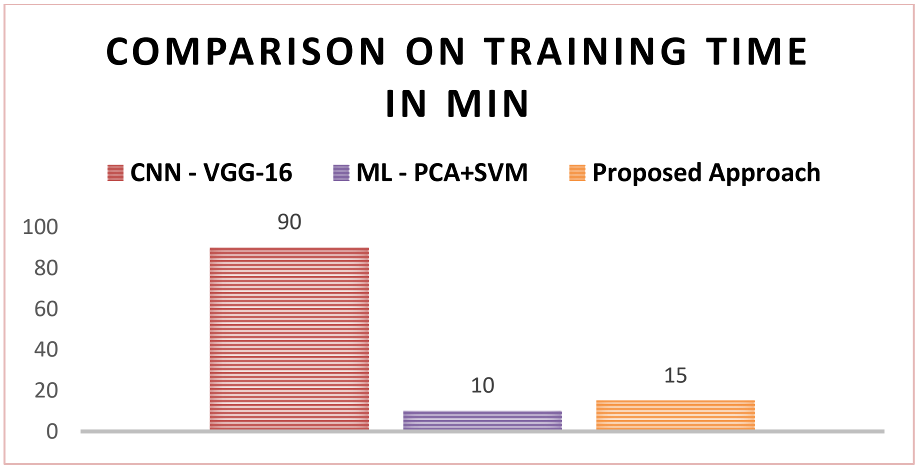


Figure 6.7: Comparison on model training time of Panel Fault Detection

6.4 Approach Taken for PV Panel Tabular Fault Classification

As a continuation of PV fault classification, we are now turned our focus on classifying tabular faults from a simulated 250 kW photovoltaic power plant. Simulated data consists with three faults type namely Class 1, 2 & 3.

Fault distribution in the dataset is as follows,

Fault Type	*Count*	*Percentage*
No Fault	100	16.67
Type 1	153	25.5
Type 2	149	24.83
Type 3	198	33

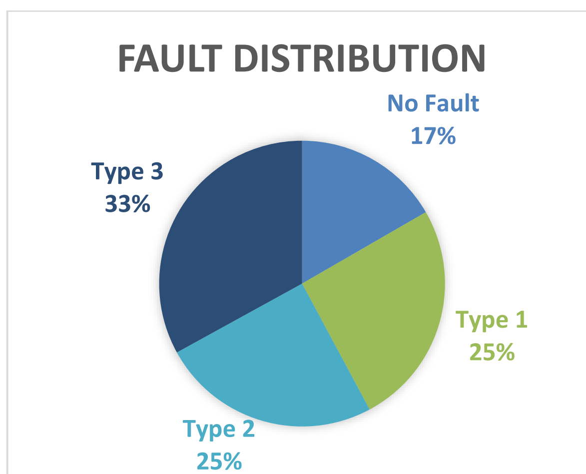


Figure 6.8: Fault distribution in Simulated power plant data

6.4.1 Exploring Different Machine Learning Based Classification Method

Based on the literature study and past experience we have seen for small amount of dataset with relatively small set of features traditional machine learning based classification method working pretty well.

Additionally, while looking at the distribution we also observed that there are no imbalance in different fault data is available hence no over sampling or under sampling steps needs to be followed for making the dataset balance.

Keeping these factors in mind we have explored below list of models for this problem,

Table 6.8: Explored Classification Model

ML & DL Classification Model		
▪ Logistic Regression	▪ Decision Tree	▪ Random Forest
▪ Support Vector Machine		

6.4.1.1 Classification Model Hyperparameter Tuning & Exploration

Hyperparameter tuning is the best way to find out the optimal parameters and corresponding best models. To find out right choice between the applied algorithm we have consider couple of deciding factors as below,

- **Classification Accuracy** – Obvious first criterion for selecting any model is accuracy of the classification which we have observed using confusion matrix and classification reports.
- **Time to Train & Inference** – As a next criterion, we have chosen the training and inferencing time of the model.
- **Nature of the model** – As a last criterion we have selected a model which is simple in nature without any programming complexities involved.

6.4.1.1.1 Tuning parameter combination

For every model tuning parameter combination is different. We will show the tuning parameters used for this study in below table,

Table 6.9: Hyperparameter Tuning explored for Tabular Fault Classification

Model Name	*Parameters & Value*	*Purpose*
Logistic Regression	C : [0.1,1,10]	Regularization parameter that controls the trade-off between fitting the data and complexity of the model.
Logistic Regression	max_iter : [50,70,100,150,180]	Maximum number of iterations taken for the solvers to converge.
Decision Tree	max_depth : [None, 10, 20, 30]	Maximum depth of the tree to control overfitting.
Random Forest	n_estimators :[10, 50, 100]	Number of trees in the forest to control the ensemble size.
Random Forest	max_depth : [None, 10, 20, 30]	Maximum depth of each tree to control overfitting.
Support Vector Machine	C : [0.1,1,10]	Regularization parameter that controls the trade-off between margin size and classification error.
Support Vector Machine	gamma: [0.1, 0.01, 0.001]	Kernel coefficient that controls the influence of individual training samples.
Every Model	Cross Validation : 5	Number of folds in cross-validation to evaluate the model performance.

Table 6.10: Hyperparameter Tuning Results

Logistic Regression

C	*max_iter*	*Accuracy*
0.1	50	0.9
0.1	70	0.91
0.1	100	0.92
0.1	150	0.93
0.1	180	0.93

Random Forest

n_estimators	*max_depth*	*Accuracy*
10	None	0.96
10	10	0.97
10	20	0.98
10	30	0.99
50	None	0.97

1	50	0.94
1	70	0.95
1	100	0.96
1	150	0.97
1	180	0.99
10	50	0.95
10	70	0.96
10	100	0.97
10	150	0.98
10	180	0.99

50	10	0.98
50	20	0.99
50	30	0.99
100	None	0.98
100	10	0.99
100	20	1
100	30	1

Decision Tree

max_depth	*Accuracy*
None	0.95
10	0.96
20	0.99
30	0.98

Support Vector Machine

C	*gamma*	*Accuracy*
0.1	0.1	0.78
0.1	0.01	0.79
0.1	0.001	0.8
1	0.1	0.81
1	0.01	0.82
1	0.001	0.86
10	0.1	0.84
10	0.01	0.85
10	0.001	0.86

6.4.1.1.2 Rationale Behind Model Selection and Associated Challenges

While exploring hyperparameter tuning, we observed that due to the limited amount of available fault data, all the models were able to converge to good results in significantly less time. As a result, classification accuracy and the time required to train and infer the results were not suitable metrics for our study. Therefore, we decided to proceed with a simpler model for our further research considering advantages during the incremental learning process.

Table 6.11: Model wise accuracy

Model Name	*Best Hyperparameter Combination*	*Accuracy*	*Best Accuracy(Hyperparameter Tuning)*
Logistic Regression	C: 1, max_iter: 180	0.95714	0.99285
Decision Tree	max_depth: 20	0.98546	0.99587
Random Forest	n_estimators: 100, max_depth: 20	0.99546	1

Support Vector Machine	C: 1, gamma: 0.001	0.78954	0.86547

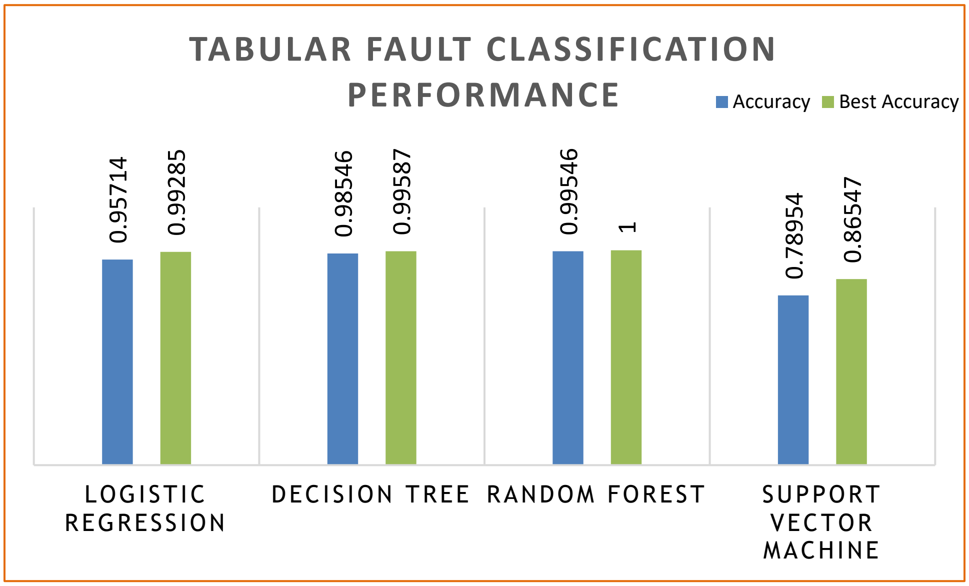


Figure 6.9: Performance comparison of different algorithm on Tabular Fault data

Based on the results in table 6.10, we have decided to choose Logistic Regression, even though its best accuracy is only the third highest among the models listed. This choice is driven by two key factors:

- **Flexibility with Larger Datasets**: Logistic Regression is highly adaptable and works efficiently with stochastic gradient descent (SGD), which is particularly beneficial when dealing with larger datasets. The ability to effectively use SGD makes it a robust choice for scenarios where data size can vary or grow over time.
- **Simplicity of Architecture**: Logistic Regression has the simplest architecture compared to the other models. This simplicity is advantageous because it reduces the complexity of implementation and maintenance. It also typically results in faster training and inference times, which are crucial for practical applications.

Considering these factors, Logistic Regression becomes a better choice for incremental learning. Its flexibility and simplicity ensure that as new data becomes available, the model can be updated efficiently without significant computational overhead. This makes it well-suited for continuous improvement and adaptation in dynamic environments.

Below is the result of Logistic regression on unseen test data,

```
Best Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        30
           1       1.00      0.97      0.99        40
           2       0.97      1.00      0.98        30
           3       1.00      1.00      1.00        40

    accuracy                           0.99       140
   macro avg       0.99      0.99      0.99       140
weighted avg       0.99      0.99      0.99       140

Logistic Regression Confusion Matrix:
[[30  0  0  0]
 [ 0 39  1  0]
 [ 0  0 30  0]
 [ 0  0  0 40]]
```

Figure 6.10: Logistic Regression Results for Fault Classification in Simulated Power Plant Data

6.5 Multimodal Approach & Incremental Learning in Fault Classification

After determining the best models for fault classification on both types of fault data, we implemented a late fusion (decision-level fusion) multimodal approach to combine these models. This method involves merging the outputs of the individual models to make a final decision. Based on the criticality level of the fault, the final classified fault result is displayed to the operator, enabling them to take appropriate action.

Below two results will show the power of incremental learning:

The comparison between the initial learning and incremental learning phases demonstrates a significant improvement in the model's predictive capabilities. Initially in fig 6.11, the model inaccurately assessed the solar panel as clean with no probable faults. However, after undergoing incremental learning, in fig 6.12 the model accurately identified electrical damage on the same solar panel, highlighting the severity and priority of the issue. This showcases the effectiveness of incremental learning in enhancing the model's ability to detect and classify faults more accurately, thereby providing more reliable and actionable insights for operators to take necessary actions.

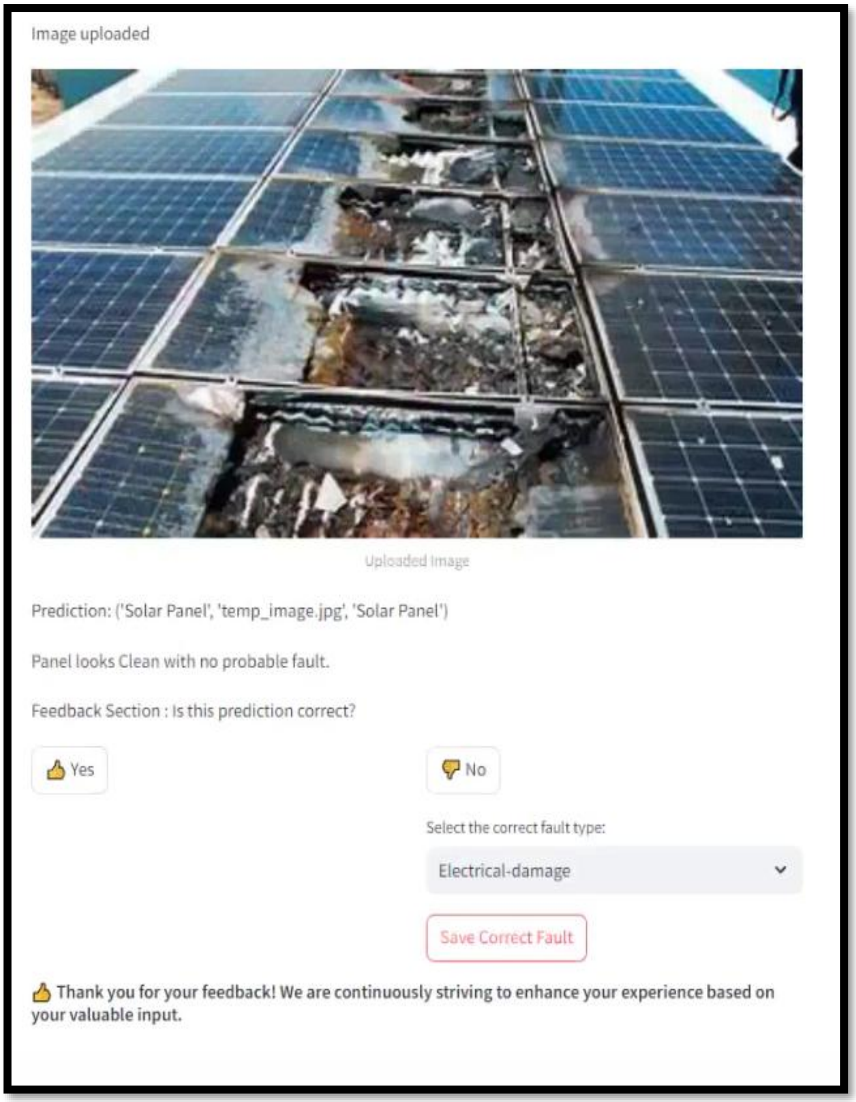


Figure 6.11: Prediction on initial learning

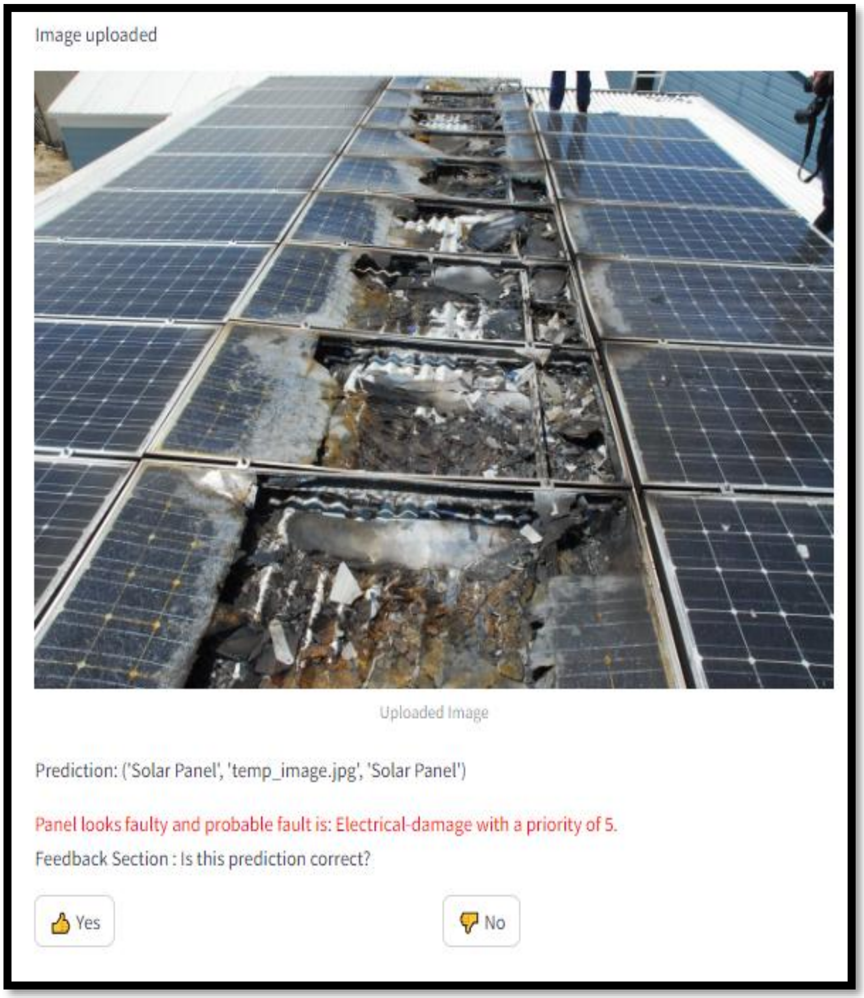


Figure 6.12: Prediction after incremental learning

Proposed approach of Fault classification now run on different dataset after incremental learning on different datasets and results looks promising across different faults as shown below,

Uploaded Image

Prediction: ('Solar Panel', 'temp_image.jpg', 'Solar Panel')

Panel looks faulty and probable fault is: Bird-drop with a priority of 1.

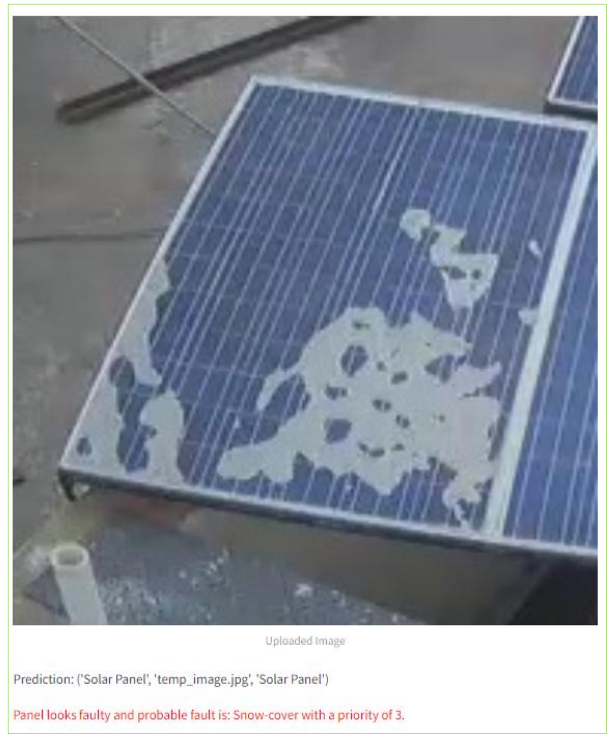
Uploaded Image

Prediction: ('Solar Panel', 'temp_image.jpg', 'Solar Panel')

Panel looks faulty and probable fault is: Dusty with a priority of 3.

Bird Drop Fault

Dusty Panel



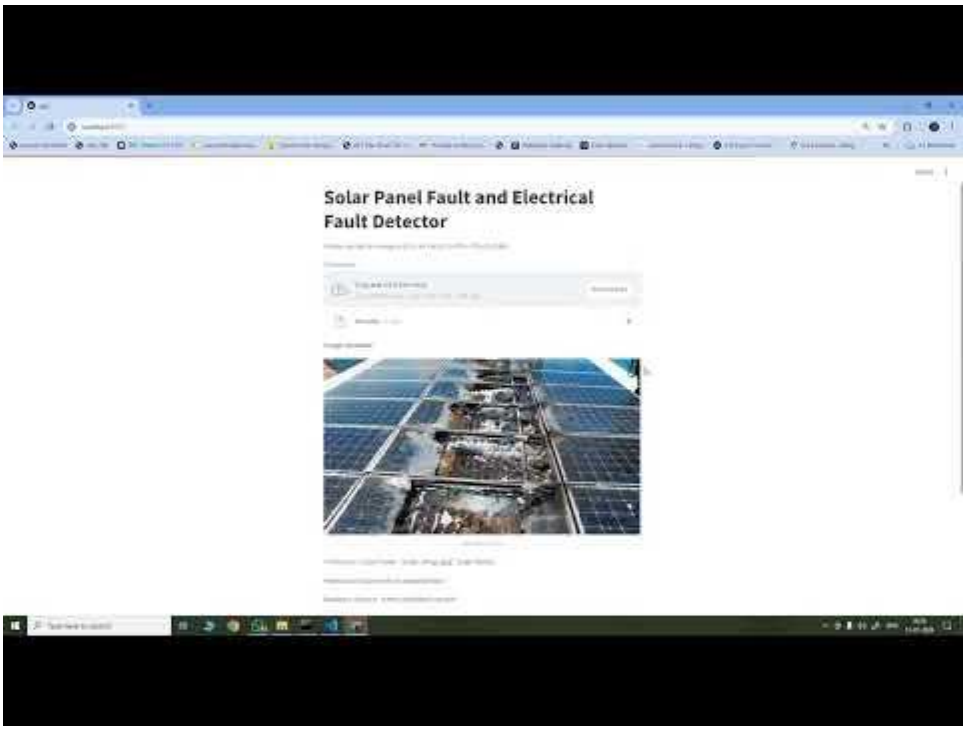
Red Dust Fault

Snow Cover

Figure 6.13: Fault classification in different dataset

6.6 Application Demonstration

Below recorded video of application is uploaded in YouTube platform to embed it in word document.



CHAPTER 7

CONCLUSION AND FUTURE WORK

This research has significantly advanced the field of predictive maintenance for solar photovoltaic (PV) systems by integrating deep learning and image analytics techniques to enhance fault detection and classification. The study's novel approach, which combines machine learning algorithms, transfer learning, and incremental learning methods, has successfully addressed the critical challenge of identifying and quantifying faults in PV panels.

7.1 Key Findings:

1. **Enhanced Predictive Maintenance**: The integration of both tabular and image faults together in classification into an existing predictive maintenance framework has demonstrated a substantial improvement in identifying and addressing PV panel faults. This comprehensive system not only predicts potential faults in panel but also identify electrical faults, enabling more effective maintenance strategies.
2. **Accuracy in Fault Detection**: The study employed a hierarchical ensemble classification system enhanced by dynamic feature extraction and dimensionality reduction, significantly improving the accuracy of fault detection. The combination of PCA, SVM, and Naive Bayes classifiers has proven effective in distinguishing various fault types and conditions in PV panels.
3. **Real-time Application and Scalability**: The developed models exhibit robust performance across diverse operational conditions, highlighting their potential for real-time application in monitoring and maintaining solar power systems. The scalability of these models ensures their adaptability to different environmental conditions and fault types, making them suitable for global deployment.
4. **Integration of Feedback Mechanisms**: The incorporation of user feedback into the classification models has enhanced the accuracy and robustness of the system. By continually updating the models with new data and corrections, the system remains adaptive and improves over time, ensuring high reliability in fault detection.

7.2 Future Work:

- **Segmentation of Fault Area:** Future work should further refine the segmentation of fault areas to provide even more precise measurements of the affected regions. By improving the accuracy of fault segmentation and calculating the percentage of the panel affected, more effective maintenance strategies can be developed. This involves using advanced segmentation algorithms and enhancing the precision of the Shoelace formula to ensure that even the smallest faults are detected and accurately measured. Accurate segmentation will enable targeted maintenance, reducing unnecessary interventions and optimizing resource allocation, ultimately leading to prolonged panel life and improved energy efficiency.
- **Expansion of Dataset and Model Training**: Future research should focus on expanding the dataset to include a wider variety of PV panel faults and operational conditions. This will further enhance the generalizability and accuracy of the models. Additionally, exploring advanced training techniques such as federated learning could improve model performance across distributed datasets.
- **Integration with IoT and Smart Grid Technologies**: Integrating the developed fault detection system with IoT devices and smart grid technologies can enable real-time monitoring and automated maintenance of PV systems. This will facilitate a more proactive approach to managing solar power installations, reducing downtime, and improving overall energy output.
- **Exploration of Advanced Deep Learning Models:** Further exploration of advanced deep learning models such as transformers and graph neural networks could offer new perspectives in fault detection and classification. These models, known for their superior handling of complex data structures, could enhance the accuracy and efficiency of the fault detection system.

REFERENCES

[1]. Yousif, Hayder, and Zahraa Al-Milaji. "Fault detection from PV images using hybrid deep learning model." *Solar Energy* 267 (2024): 112207.

[2]. Kaitouni, Samir Idrissi, Ibtihal Ait Abdelmoula, Niima Es-sakali, Mohamed Oualid Mghazli, Houda Er-retby, Zineb Zoubir, Fouad El Mansouri, Mohammed Ahachad, and Jamal Brigui. "Implementing a Digital Twin-based fault detection and diagnosis approach for optimal operation and maintenance of urban distributed solar photovoltaics." *Renewable Energy Focus* 48 (2024): 100530.

[3]. Duranay, Zeynep Bala. "Fault Detection in Solar Energy Systems: A Deep Learning Approach." *Electronics* 12.21 (2023): 4397.

[4]. Mustafa, Zain, Ahmed SA Awad, Maher Azzouz, and Ahmed Azab. "Fault identification for photovoltaic systems using a multi-output deep learning approach." *Expert Systems with Applications* 211 (2023): 118551.

[5]. Constantin, Alexandru-Ionel, et al. "Importance of Preventive Maintenance in Solar Energy Systems and Fault Detection for Solar Panels based on Thermal Images." *Electrotehnica, Electronica, Automatica* 71.1 (2023).

[6]. Eldeghady, Ghada Shaban, Hanan Ahmed Kamal, and Mohamed A. Moustafa Hassan. "Fault diagnosis for PV system using a deep learning optimized via PSO heuristic combination technique." *Electrical Engineering* (2023): 1-15.

[7]. Hassan, Sharmarke, and Mahmoud Dhimish. "A Survey of CNN-Based Approaches for Crack Detection in Solar PV Modules: Current Trends and Future Directions." *Solar*. Vol. 3. No. 4. MDPI, 2023.

[8]. Van Gompel, Jonas, Domenico Spina, and Chris Develder. "Cost-effective fault diagnosis of nearby photovoltaic systems using graph neural networks." *Energy* 266 (2023): 126444.

[9]. Pathak, Sujata P., Sonali Patil, and Shailee Patel. "Solar panel hotspot localization and fault classification using deep learning approach." *Procedia Computer Science* 204 (2022): 698-705.

[10]. Haidari, Parsa, et al. "Deep learning-based model for fault classification in solar modules using infrared images." *Sustainable Energy Technologies and Assessments* 52 (2022): 102110.

[11]. Kumar Chawrasia, Sandeep, Debmalya Hembram, Dipanjan Bose, and Chandan Kumar Chanda. "Deep learning assisted solar forecasting for battery swapping stations." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 46, no. 1 (2024): 3381-3402.

[12]. Maciel, Joylan Nunes, Jorge Javier Gimenez Ledesma, and Oswaldo Hideo Ando Junior. "Hybrid prediction method of solar irradiance applied to short-term photovoltaic energy generation." *Renewable and Sustainable Energy Reviews* 192 (2024): 114185.

[13]. Miraftabzadeh, Seyed Mahdi, Cristian Giovanni Colombo, Michela Longo, and Federica Foiadelli. "A day-ahead photovoltaic power prediction via transfer learning and deep neural networks." *Forecasting* 5, no. 1 (2023): 213-228.

[14]. Seghiour, Abdellatif, Hamou Ait Abbas, Aissa Chouder, and Abdlhamid Rabhi. "Deep learning method based on autoencoder neural network applied to faults detection and diagnosis of photovoltaic system." *Simulation Modelling Practice and Theory* 123 (2023): 102704.

[15]. Khan, Waqas, Shalika Walker, and Wim Zeiler. "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach." *Energy* 240 (2022): 122812.

[16]. Venkatesh, S. Naveen, and V. Sugumaran. "Fault Detection in aerial images of photovoltaic modules based on Deep learning." *IOP Conference Series: Materials Science and Engineering*. Vol. 1012. No. 1. IOP Publishing, 2021.

[17]. Abubakar, Ahmad, Carlos Frederico Meschini Almeida, and Matheus Gemignani. "Review of artificial intelligence-based failure detection and diagnosis methods for solar photovoltaic systems." *Machines* 9.12 (2021): 328.

[18]. Pierdicca, Roberto, et al. "Automatic faults detection of photovoltaic farms: solAIr, a deep learning-based system for thermal images." *Energies* 13.24 (2020): 6496.

[19]. Kim, Gyu Gwang, et al. "Fault detection for photovoltaic systems using multivariate analysis with electrical and environmental variables." *IEEE Journal of Photovoltaics* 11.1 (2020): 202-212.

[20]. Zyout, Imad, and Abdulrohman Oatawneh. "Detection of PV solar panel surface defects using transfer learning of the deep convolutional neural networks." *2020 Advances in Science and Engineering Technology International Conferences (ASET)*. IEEE, 2020.

[21]. Jingwei Zhang et al. "A fault severity quantification approach of photovoltaic array based on pre-estimation and fine-tuning of fault parameters" 2023 Journal of Renewable & Sustainable Energy https://doi.org/10.1063/5.0152868

[22]. Shihavuddin, A., et al. (2017). "Deep Convolutional Neural Networks for Automated Detection of Defects in Solar Panels." Journal of Imaging, 3(1), 24.

[23]. Pierdicca, R., et al. (2018). "Image Pattern Recognition in Photovoltaic Systems: A Literature Review." *Remote Sensing*, 10(8), 1234.

[24]. Amaral, S., et al. (2019). "Fault Diagnosis in PV Tracking Systems Using Image Processing and Principal Component Analysis." *IEEE Transactions on Industrial Electronics*, 66(11), 8792-8802.

[25]. Zhang, et al. (2022). "Fault Detection in Solar Energy Systems: A Deep Learning Approach." *Electronics Journal*.