Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

Hybrid Transformer-Based Solar PV Panel Detection and Energy Estimation System

[1]J. S. S. Prasanna, [2]S. Chandra Sekhar

[1]PG Scholar, Department of Computer Science and Engineering, University College of Engineering Kakinada (A), JNTUK, India
[2]Assistant Professor, Department of Computer Science and Engineering, University College of Engineering Kakinada (A), JNTUK, India

Abstract: The increasing installation of rooftop solar photovoltaic (PV) systems has created a need for automated techniques that can identify photovoltaic panels from aerial and satellite imagery and convert image-based detections into useful solar-energy information. Manual rooftop inspection is time-consuming and can be inconsistent when panels are small, closely arranged, partially obscured, or surrounded by visually similar structures. This paper presents a Hybrid Transformer-Based Solar PV Panel Detection and Energy Estimation System that combines convolutional feature extraction with Transformer-based contextual learning for pixel-level photovoltaic segmentation. A pretrained ResNet-18 backbone extracts hierarchical spatial features from rooftop imagery, while a Transformer Encoder models relationships among spatially separated image regions. A transposed-convolution decoder reconstructs the representation into a binary segmentation mask. The predicted mask is subsequently processed using contour extraction to obtain panel-level bounding boxes. The segmented area is converted into physical area using image spatial resolution and is then used to estimate installed photovoltaic capacity and expected daily electricity generation. The complete workflow is integrated into a Streamlit application for interactive image upload, detection visualization, panel localization, performance evaluation, and solar analysis. The study demonstrates that combining local CNN representations with global contextual modelling provides an effective basis for integrated rooftop PV detection, localization, and solar-energy assessment.

Keywords: Solar Photovoltaic Detection, Rooftop Imagery, Hybrid CNN-Transformer, ResNet-18, Transformer Encoder, Semantic Segmentation, Solar Energy Estimation.

I. INTRODUCTION

The transition toward distributed renewable energy has increased the importance of understanding where photovoltaic installations are located and how much generation potential they may represent. Rooftops are particularly attractive for decentralized solar deployment because they can provide generation space without requiring additional land. However, large-scale assessment of rooftop installations from aerial or satellite imagery is difficult when performed manually. An analyst must inspect many images, identify photovoltaic regions, determine their boundaries, and interpret the resulting installation size. Such a workflow is expensive in time and can produce inconsistent estimates when imagery contains shadows, roof structures, vegetation, different panel orientations, or low-contrast surfaces.

Computer vision provides a practical alternative by treating rooftop PV identification as an image-analysis problem. Object detection can localize photovoltaic modules using bounding boxes, whereas semantic segmentation can classify individual pixels and therefore represent the actual spatial footprint of a panel region more precisely. For solar-resource assessment, this pixel-level representation is valuable because panel area can be derived from the segmentation mask and subsequently related to installed capacity and energy generation. The literature reviewed in the thesis includes deep-learning approaches for photovoltaic module segmentation, aerial-image detection, remote-sensing analysis, and solar-energy assessment.

A limitation of purely convolutional approaches is that their strongest representations are learned from local receptive fields. Rooftop scenes, however, contain spatial relationships that may extend over larger regions. Panels can occur in groups, align with roof geometry, or appear next to structures with similar texture and reflectance. Transformer architectures complement convolutional feature learning by using attention to model relationships among different feature tokens. The proposed work therefore combines a ResNet-18 CNN backbone with a Transformer Encoder, using the CNN to capture detailed local structure and the Transformer to incorporate broader contextual

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

information. The hybrid representation is then decoded into a binary PV segmentation mask.

The contribution of this work is not limited to segmentation. The predicted mask is converted into object-level regions through contour extraction, allowing bounding boxes and panel coordinates to be generated. The system then estimates physical panel area using the spatial resolution of the image and derives approximate installed capacity and daily electricity generation using predefined photovoltaic parameters. A Streamlit interface brings these functions together so that a user can upload an image and inspect the input, segmentation, localized panels, evaluation metrics, and solar-analysis outputs in a single environment. This integrated design addresses a functional gap between image-based PV detection and practical solar assessment.

An integrated approach for rooftop photovoltaic analysis, combining image acquisition, preprocessing, deep feature extraction, Transformer-based contextual learning, semantic segmentation, panel localization, area estimation, and solar-energy prediction. The system is further implemented through a Stream lit-based interactive application, allowing the detection results, segmentation masks, performance measures, panel-level information, and solar-energy estimates to be visualized through a single interface. Overall, the approach aims to provide a more automated and scalable solution for rooftop solar PV assessment and can support applications related to renewable-energy monitoring and solar-resource planning.

II. LITERATURE REVIEW

Research in rooftop photovoltaic detection has increasingly moved from conventional image processing toward deep-learning-based segmentation and detection. Study [1] describes a deep-learning approach for photovoltaic module segmentation in high-resolution aerial and satellite imagery and emphasizes the value of Transformer-driven semantic segmentation and multi-scale feature learning. The work highlights the challenge of maintaining reliable performance when image resolution and environmental conditions change. These observations motivate architectures that can represent both detailed panel appearance and broader scene context.

A rooftop PV extraction framework using aerial imagery and building-footprint information is described in [2]. The use of auxiliary building context can suppress false detections in complicated urban scenes, but dependence on external footprint data may limit portability when equivalent geographic information is unavailable. Study [3] examines the reliability and transferability of deep-learning models for rooftop PV detection across space and emphasizes that performance can vary with geographic and imaging conditions. Such findings reinforce the importance of robust representations and evaluation under controlled conditions.

Review-oriented studies in remote sensing and artificial intelligence [4], [6], [15] describe the rapid development of deep learning for image understanding and emphasize issues such as domain variation, computational demand, and generalization. Survey work on photovoltaic panel detection [5] identifies image processing, object detection and segmentation as important directions. Dataset-oriented work such as [7] demonstrates the increasing availability of large-scale photovoltaic imagery for training detection systems, while [8] connects rooftop PV detection with energy assessment. Comparative model studies [9] further indicate that model selection can affect detection quality under aerial-image conditions.

Other studies focus on monitoring and analysis after detection. Work on rooftop PV development monitoring [10] demonstrates the value of repeated aerial imagery for understanding deployment patterns. Research using three-dimensional remote-sensing information [11] and UAV thermal imagery [12] illustrates how additional sensing modalities can support photovoltaic localization and labelling. Studies [16] and [17] extend deep-learning-based analysis toward solar potential assessment, while [18] reviews object-detection approaches in remote-sensing imagery. CNN-based rooftop detection and vision-based solar-panel monitoring are also investigated in [19] and related work.

The reviewed literature indicates three recurring requirements. First, PV systems must be localized accurately despite complex rooftop backgrounds. Second, segmentation should preserve spatial boundaries sufficiently well for quantitative area measurement. Third, an operational system should translate visual detections into useful downstream information rather than stopping at a binary mask. The proposed Hybrid CNN-Transformer framework addresses these requirements by combining local convolutional features, attention-based contextual modelling, pixel-level segmentation, contour-based localization, spatial quantification, and energy estimation.

Table 1 summarizes the relationship between representative directions in the reviewed literature and the focus of the proposed system.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

Table 1: Research Gaps from Literature Survey

Research direction	Primary strength	Gap addressed by proposed system
Transformer / multi-scale PV segmentation [1]	Global context and multi-scale representations	Integrates contextual modelling with a practical end-to-end rooftop workflow
Aerial imagery + building footprints [2]	Uses additional geographic context	Reduces dependence on auxiliary footprint information
Cross-location reliability [3]	Highlights generalization challenges	Uses hybrid local/global representation and controlled evaluation
PV detection surveys [5], [9]	Compare detection and segmentation strategies	Extends detection to area, capacity and energy estimation
Monitoring and energy assessment [8], [10], [17]	Connects PV detection with deployment or energy analysis	Provides these functions in one interactive application

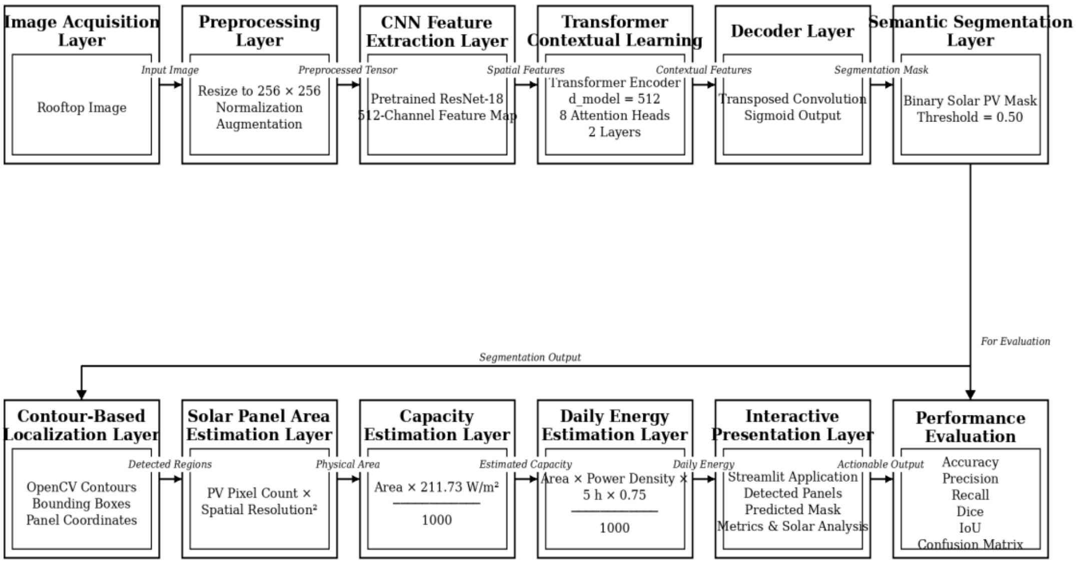


Figure 1: Architecture of the proposed Hybrid Transformer-Based Solar PV Detection System

III. PROPOSED SYSTEM

The proposed system is designed as an end-to-end pipeline for rooftop photovoltaic detection and solar-energy assessment. The architecture as represented in fig 1 contains image acquisition and preprocessing, CNN feature extraction, Transformer contextual modelling, semantic segmentation, contour-based localization, area estimation, power estimation, performance evaluation, and interactive visualization. Each stage produces an output that becomes the input to the next stage, allowing the system to transform an aerial rooftop image into both visual and quantitative information.

3.1 Image Acquisition and Preprocessing

During model development, rooftop imagery is obtained from the PV01 Rooftop Brick Dataset described in the thesis. Corresponding binary label images are used as reference masks. The preprocessing stage establishes a consistent representation before deep learning. Images are resized to 256 × 256 pixels, normalized to a suitable numerical range, and converted into tensor representations compatible with PyTorch. The training workflow also supports controlled augmentation to increase variation in the visual appearance of rooftops. These operations reduce differences caused by image dimensions and improve the consistency of feature learning.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

The use of a binary segmentation target simplifies the learning objective: each pixel belongs either to the photovoltaic-panel class or to the background class. This formulation is appropriate for estimating the spatial footprint of PV installations. During inference, the same image preparation pipeline is applied before the trained model produces the segmentation prediction. The application additionally allows the user to specify spatial resolution so that pixel measurements can be converted into physical area.

3.2 ResNet-18 Feature Extraction

The first deep-learning stage uses a pretrained ResNet-18 backbone. ResNet-18 is used as a compact convolutional feature extractor capable of learning hierarchical visual representations. Early layers capture low-level information such as edges and textures, while deeper layers encode higher-level structural characteristics. In rooftop imagery, these representations can describe panel edges, repeated module patterns, roof boundaries, and other visual structures that help separate photovoltaic surfaces from the surrounding scene.

In the implemented model, the final classification layers of ResNet-18 are removed and the convolutional feature maps are retained. For an input image, the resulting representation has 512 channels. This representation preserves spatial structure while providing a rich feature space for the subsequent Transformer stage. The use of a pretrained backbone also provides an effective initialization for learning visual patterns from the available rooftop dataset.

3.3 Transformer Encoder for Contextual Learning

The CNN feature map is reshaped into a sequence of feature tokens. Each spatial location becomes a token and the resulting sequence is processed by a Transformer Encoder. The implemented Transformer uses an embedding dimension of 512, eight attention heads, and two encoder layers. Self-attention enables each token to incorporate information from other spatial locations rather than depending only on neighbouring convolutional receptive fields.

This contextual operation is particularly relevant to rooftop PV imagery. A panel may have a visual appearance similar to another rooftop material, and local texture alone may be insufficient for classification. Information from nearby panels, roof layout, orientation, and larger contextual structures can provide additional evidence. The Transformer therefore complements rather than replaces the CNN: convolution provides detailed local structure, while attention provides relationships among spatially separated regions.

3.4 Decoder and Semantic Segmentation

After Transformer processing, the token sequence is restored to its spatial feature-map arrangement and passed to a decoder composed of transposed convolution layers. The decoder progressively increases spatial resolution and ultimately produces a single-channel output with a sigmoid activation. The output represents the probability that each pixel belongs to a photovoltaic panel. A threshold of 0.50 is used in the implemented prediction workflow to convert the probability map into a binary mask.

The segmentation output is the central representation of the system because it preserves the shape and occupied area of detected PV regions. Unlike a bounding-box-only detector, a segmentation mask can follow irregular boundaries and can be used directly for pixel counting. The resulting mask also provides a basis for the subsequent contour extraction, physical-area calculation, and solar-energy estimation stages.

3.5 Contour-Based Panel Localization

The binary segmentation mask is processed with OpenCV to extract connected contours. Each relevant contour represents a detected photovoltaic region in the predicted mask. Bounding rectangles are calculated from the contours, producing top-left, top-right, bottom-left, and bottom-right coordinate information. This step converts pixel-level segmentation into an object-level representation and makes it possible to analyse detected regions individually.

The contour-based approach is intentionally lightweight because the segmentation network has already performed the main visual discrimination. OpenCV is therefore used as a post-processing layer rather than as a separate learned detector. The resulting bounding boxes are displayed over the input image in the application, allowing users to verify the spatial locations of detected panels.

3.6 Solar Panel Area Estimation

Once the PV pixels have been identified, their count is converted into physical area using the spatial resolution of the input imagery. If N is the number of segmented PV pixels and r is the spatial resolution in metres per pixel, the estimated panel area A is given by Equation (1):

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

A = N × r²…….. (1)

In the implemented application, the default spatial resolution is 0.15 m/pixel, although the user interface allows this value to be adjusted. The area calculation provides a direct quantitative bridge between image segmentation and solar-resource analysis. Panel-wise areas can also be calculated from individual connected regions, enabling comparison among detected installations.

3.7 Capacity and Daily Energy Estimation

The estimated panel area is converted into an approximate installed capacity using the effective photovoltaic power density adopted by the project. With P_d denoting power density, the capacity C in kilowatts is expressed by Equation 2 as:

C = (A × P_ d) / 1000…….. (2)

The implemented value for effective photovoltaic power density is 211.73 W/m². Daily electricity generation is then estimated using average peak sun hours H and overall system efficiency η is expressed by Equation (3) as:

E_daily = A × P_d × H × η / 1000….. (3)

The project uses H = 5 hours/day and η = 0.75. These values are predefined engineering assumptions rather than direct measurements of instantaneous irradiance or operating conditions. Consequently, the resulting energy figure should be interpreted as an approximate potential rather than a site-specific production forecast.

3.8 Interactive Streamlit Application

The complete pipeline is integrated into a Streamlit-based application. The interface allows a user to upload a rooftop image, specify spatial resolution and solar parameters, and view the resulting analysis. The output includes the original image, detected panels with bounding boxes, the predicted segmentation mask, panel coordinates, estimated area, capacity and daily energy generation, and model-performance information as represented in Fig2. This integration reduces the need for direct interaction with the underlying Python implementation and provides a practical demonstration of the research workflow.

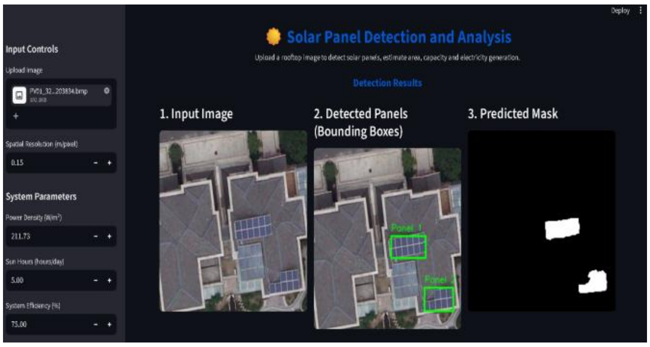


Figure 2: Streamlit user interface for rooftop PV detection and solar analysis

IV. METHODOLOGY AND EXPERIMENTAL SETUP

The experimental methodology follows the same sequential structure used in the implemented system: image preparation, hybrid feature extraction, segmentation, localization, spatial quantification, energy estimation, and evaluation. The evaluation procedure is controlled so that dataset selection, preprocessing, model configuration, testing conditions, spatial resolution, and solar-analysis parameters remain consistent. This reduces the influence of unrelated variations and makes the reported results representative of the implemented model.

4.1 Training Configuration

The training implementation uses PyTorch and a DataLoader with a batch size of four. The network is trained for 80 epochs using the Adam optimizer with a learning rate of 0.0001. Binary Cross-Entropy (BCELoss) is used as the training loss for the single-channel sigmoid segmentation output. Model weights are saved in PyTorch format and subsequently loaded by the prediction and Streamlit application components.

The architecture uses the pretrained ResNet-18 convolutional encoder followed by a two-layer Transformer Encoder with d_model = 512 and eight attention heads. The decoder contains transposed convolution operations that progressively reconstruct the segmentation representation. Input images are resized to 256 × 256 pixels, and the prediction threshold is set to 0.50 in the application. These implementation settings provide a reproducible description of the developed prototype.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

Table 2: Software and Hardware Environment

Component	Technology / Requirement	Purpose
Programming	Python 3.9+	Core implementation
Deep learning	PyTorch	Model construction, training and inference
CNN backbone	ResNet-18	Hierarchical spatial feature extraction
Context modelling	Transformer Encoder	Long-range spatial relationships
Computer vision	OpenCV	Contours, localization and image processing
Data / numerics	NumPy, Pandas	Data handling and calculations
Augmentation	Albumentations	Image preprocessing and augmentation
Evaluation	Scikit-learn	Performance statistics and confusion matrix
Visualization	Matplotlib	Training and performance graphs
Web application	Streamlit	Interactive prediction and visualization
Model storage	.pth	Saving and loading trained weights

4.2 Evaluation Metrics

The model is evaluated using Accuracy, Precision, Recall, Dice Coefficient, and Intersection over Union (IoU). These metrics capture complementary aspects of pixel-level classification and spatial segmentation quality. Accuracy measures the proportion of correctly classified pixels across the image as represented in Equation 4.

Accuracy = (TP + TN) / (TP + TN + FP + FN)……… (4)

Precision measures the fraction of predicted PV pixels that are actually PV pixels as represented in Equation 5.

Precision = TP / (TP + FP)……… (5)

Recall measures the fraction of actual PV pixels successfully recovered by the model as represented in Equation 6.

Recall = TP / (TP + FN)………(6)

Dice measures the overlap between predicted and reference PV regions as expressed in Equation 7.

Dice = 2TP / (2TP + FP + FN)…….. (7)

IoU uses the union of the two regions and is therefore expressed in equation 8. A pixel-level confusion matrix is also generated to expose true-positive, true-negative, false-positive, and false-negative classifications.

IoU = TP / (TP + FP + FN) ……… (8)

Together, these measures provide a more complete assessment than any single score.

4.3 Overall Workflow

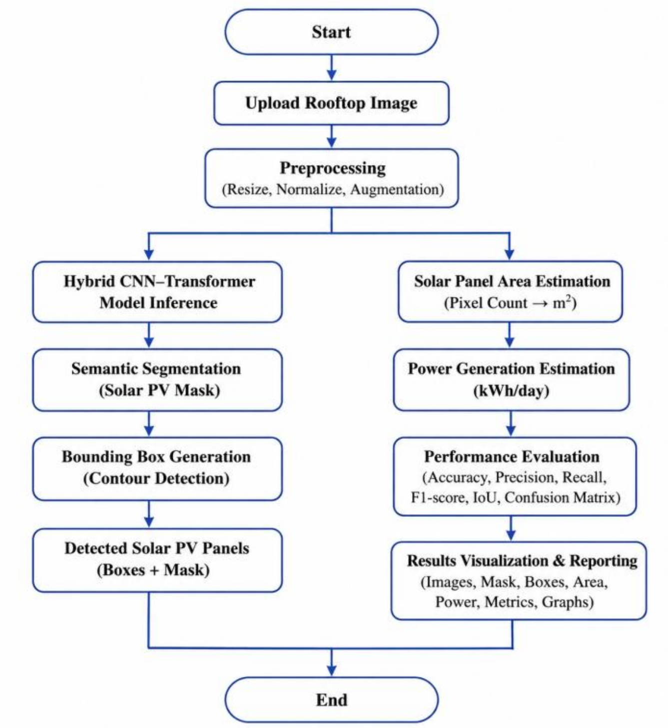


Figure 3: End-to-end execution workflow of the proposed system

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

From the fig 3 the end-to-end executional workflow is been demonstrated the operational sequence begins with image upload or dataset acquisition. The image is resized and normalized, after which the CNN extracts hierarchical spatial features. The feature map is converted to a token sequence and processed by the Transformer Encoder. The resulting contextual representation is reshaped and decoded into a binary segmentation mask. OpenCV contour processing converts connected PV regions into bounding boxes. Segmented pixels are counted and transformed into physical area, which is then used for capacity and daily-energy estimation. Finally, the application displays the visual and quantitative outputs and calculates the evaluation metrics when reference information is available.

V. RESULTS

5.1 Detection and Segmentation Output

The developed system successfully produces three principal visual outputs for an uploaded rooftop image: the original input, the detected-panel image with bounding boxes, and the predicted binary segmentation mask. The example shown in Fig. 4 contains two visible photovoltaic regions. The detected regions are represented by bounding boxes, while the segmentation mask isolates the PV pixels from the surrounding rooftop background. This confirms the intended progression from image-level input to pixel-level segmentation and object-level localization.

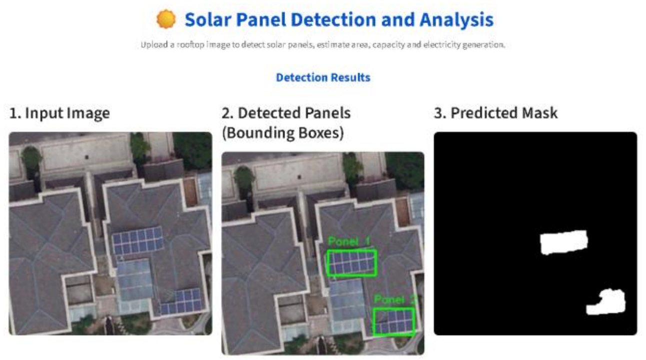


Figure 4: Example rooftop image, detected PV regions and predicted segmentation mask

The segmentation output is particularly important because the downstream area and energy calculations are derived from the detected PV pixels. The visual result indicates that the predicted regions are spatially concentrated around the visible photovoltaic installations. The contour-processing stage then converts these regions into localized panel objects, making it possible to associate coordinates and area measurements with individual detections.

5.2 Test-Image Performance

For the reported test image from Table 3, the system displays an IoU of 0.841, a Dice score of 0.914, precision of 0.982, and recall of 0.855. These values from table 3 indicate strong agreement between the predicted PV region and the reference region while also showing that the model is highly selective when it labels a pixel as photovoltaic. The difference between precision and recall is informative: the model produces relatively few false-positive PV pixels, while some reference PV pixels are not recovered. The IoU and Dice values nevertheless indicate substantial spatial overlap between prediction and reference.

Table 3: Performance metrics

Metric	Test-image score
IoU	0.841
Dice Score	0.914
Precision	0.982
Recall	0.855
Accuracy	0.993

5.3 Training Behaviour

The training curves show progressive improvement over the 80 training epochs. The Dice curve from figure 5, begins at approximately 0.72 and increases rapidly during the early epochs. It remains predominantly above 0.90 during the middle and later stages and approaches approximately 0.97 toward the final epochs. The IoU curve from fig 6also improves rapidly during early training, reaching approximately 0.70 around epoch 10, approximately 0.85 around epochs 15–18, and approximately 0.90–0.92 around epochs 35–40. It approaches approximately 0.95 during the later epochs.

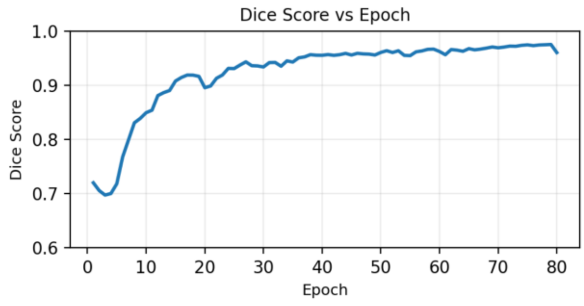


Figure 5: Training Dice Score performance across epochs

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

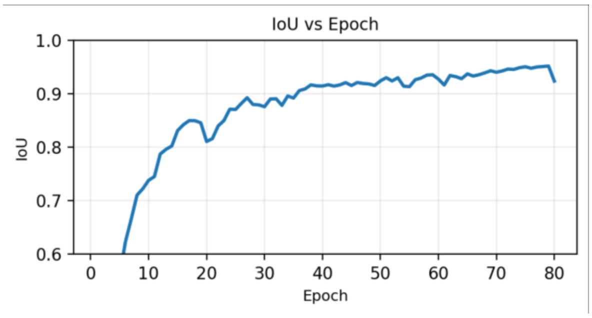


Figure 6: Training IoU performance across epochs

Precision curve from figure 7 begins near 0.78 and rises to approximately 0.83–0.86 within the first ten epochs. It reaches approximately 0.90–0.91 around epochs 10–20 and approaches 0.96–0.97 after epoch 50. The later curve is comparatively stable, indicating that the learned representation becomes increasingly discriminative for separating PV regions from similar rooftop structures.

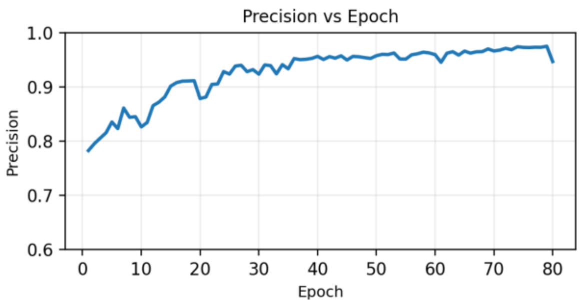


Figure 7: Training Precision performance across epochs

Recall graph from fig 8starts near 0.65 and improves substantially as training progresses. After a small temporary decline around epoch 20, the curve recovers and reaches approximately 0.92–0.96 between epochs 20 and 40. Beyond epoch 40, recall remains largely above 0.95 and reaches approximately 0.97–0.98 toward the final training stage. The overall trend indicates improved recovery of actual PV regions during optimization.

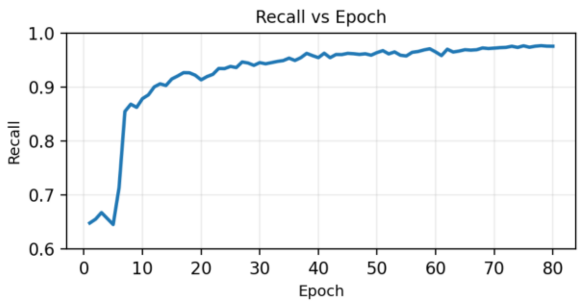


Figure 8: Training Recall performance across epochs

5.4 Solar-Energy Analysis

The system extends segmentation into quantitative solar analysis. For the example image from figure 9, two panel regions are identified. The first panel has an estimated area of 30.52 m² and the second has an estimated area of 23.31 m², giving a combined detected area of 53.83 m². Using the adopted power density of 211.73 W/m², these areas correspond to approximate installed capacities of 6.46 kW and 4.94 kW, respectively. The combined capacity is approximately 11.40 kW.

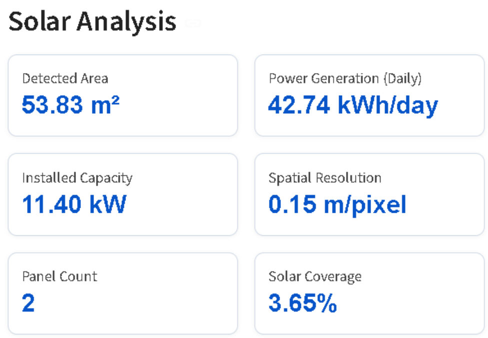


Figure 9: Solar analysis output for the detected rooftop panels

Using five peak-sun hours per day and a system efficiency of 75%, the first detected region corresponds to approximately 24.23 kWh/day and the second to approximately 18.51 kWh/day. The combined estimate is approximately 42.74 kWh/day. These values are internally consistent with the area-based formulation used by the application. They should be interpreted as approximate generation potential because the model does not directly incorporate changing irradiance, weather, tilt, shading, temperature, or seasonal conditions.

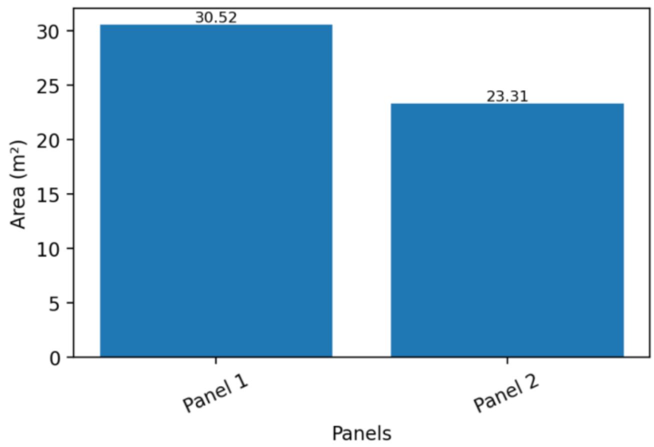


Figure 10: Panel-wise area distribution

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

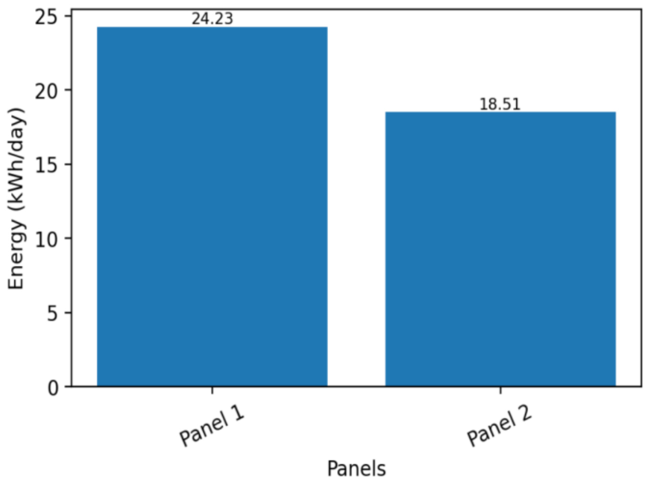


Figure 11: Panel-wise daily energy distribution

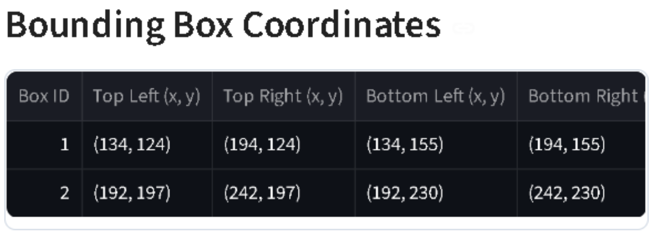
The panel-wise plot from figure 11 demonstrates the area occupied by each panel in a bar chart whereas from figure 12 we can demonstrate the daily energy distribution of the solar panels. A larger detected region contributes a larger physical area and therefore a larger estimated capacity and daily energy value under the same power-density, peak-sun-hour, and efficiency assumptions. This provides a transparent link between the segmentation result and the solar-analysis output. The table 4 illustrates the performance of the solar panels in terms of capacity, energy through area.

Table 4: Performance of the solar panels

Panel	Area (m^2)	Capacity (kW)	Energy (kWh/day)
Panel 1	30.52	6.46	24.23
Panel 2	23.31	4.94	18.51
Total	53.83	11.40	42.74

5.5 Bounding-Box and Coordinate Analysis

The system also produces coordinate information for each detected region. Bounding-box coordinates provide a compact description of panel location and extent and can be used for subsequent mapping or monitoring workflows. From the figure 12 the coordinate representation is generated after contour extraction and therefore corresponds directly to the connected regions identified by the segmentation mask. This provides object-level information without requiring a separate learned object detector in the final inference pipeline.


Bounding Box Coordinates

Box ID	Top Left (x, y)	Top Right (x, y)	Bottom Left (x, y)	Bottom Right
1	(134, 124)	(194, 124)	(134, 155)	(194, 155)
2	(192, 197)	(242, 197)	(192, 230)	(242, 230)

Figure 12: Bounding-box coordinate information for detected panels

5.6 Pixel-Level Confusion Matrix

The pixel-level confusion matrix in the experimental example from figure 13,contains 2,497 true-positive pixels, 425 false-negative pixels, 46 false-positive pixels, and 62,568 true-negative pixels. From these values, the derived accuracy is approximately 0.993, precision is approximately 0.982, recall is approximately 0.855, Dice is approximately 0.914, and IoU is approximately 0.841. The agreement between the directly displayed precision/recall values and the confusion-matrix-derived values provides an internal consistency check on the reported test-image performance.

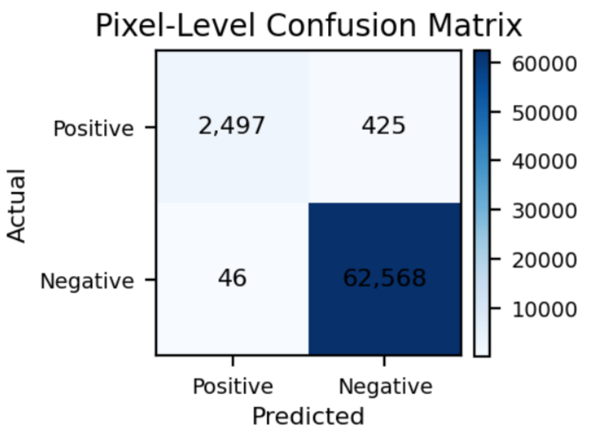


Figure 13: Pixel-level confusion matrix

The relatively small number of false-positive pixels indicates that the model is conservative when assigning the PV class, which is consistent with the high precision. The larger number of false-negative pixels explains the lower recall relative to precision. For rooftop solar applications, this distinction is important because under-segmentation can reduce estimated panel area, whereas over-segmentation can inflate the calculated solar potential. A balanced interpretation therefore requires considering precision, recall, Dice, and IoU together.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

5.7 Consolidated Results

The results can be viewed at three levels. At the visual level, the model identifies coherent PV regions and provides localized bounding boxes. At the segmentation level, the test image achieves strong overlap metrics, with IoU 0.841 and Dice 0.914. At the application level, the detected pixels are transformed into panel areas, capacities, and energy estimates. This multi-level output is a key characteristic of the proposed system because it connects computer-vision performance with a practical renewable-energy assessment workflow.

The training curves further indicate that the model learns useful representations progressively. Dice, IoU, precision, and recall all improve substantially during the training process, with relatively stable behaviour in the later epochs. Although the thesis reports a single example test-image result rather than a broad statistical benchmark across multiple independent datasets, the available results demonstrate the intended functionality and provide evidence that the hybrid architecture can produce useful rooftop PV segmentation and analysis outputs.

VI. DISCUSSION

The central design choice in the proposed system is the combination of convolutional and Transformer-based feature learning. CNNs are effective at representing local patterns that are important for PV boundaries, module textures, and rooftop structures. Transformers complement these representations by allowing spatially separated regions to influence one another through self-attention. This combination is particularly appropriate for aerial rooftop imagery because PV installations often form structured groups and can be difficult to distinguish from visually similar surfaces using local evidence alone.

The reported precision of 0.982 suggests that the model produces relatively few false-positive PV classifications for the evaluated example. This is valuable for solar assessment because false positives can lead to an overestimation of panel area and consequently overestimate installed capacity and energy potential. At the same time, recall of 0.855 indicates that some PV pixels are missed. Such missed regions can cause underestimation. The IoU and Dice scores provide a more direct view of the spatial compromise between these two effects and show that the predicted segmentation still has substantial agreement with the reference region.

Another strength is the integration of post-processing and solar analysis. A conventional segmentation model may end after generating a binary mask. In the proposed system, the mask is converted into contours and bounding boxes, then into physical area and energy estimates. This creates a more complete workflow for users who need not only to know whether a rooftop contains PV panels but also to understand the approximate size and generation potential of the detected installation.

The Streamlit implementation further improves accessibility. The user does not need to run separate preprocessing, inference, contour extraction, and calculation scripts. Instead, these stages are presented through a single interface. The application displays the original image, detected regions, mask, coordinates, metrics, and solar analysis. This makes the prototype suitable for demonstration and exploratory assessment and provides a foundation for future integration with geographic mapping, reporting, and monitoring tools.

However, several limitations should be recognized. The current energy calculation uses fixed values for effective power density, peak sun hours, and efficiency. Actual PV generation depends on irradiance, weather, season, orientation, tilt, shading, temperature, inverter behaviour, and module characteristics. The current area calculation also depends on the spatial-resolution value supplied to the application. Therefore, the numerical energy output is best treated as an approximate potential estimate rather than a calibrated production forecast.

The available experimental evidence is also limited to the outputs documented in the thesis. The reported performance values are explicitly available for a test-image example, while the training curves show the evolution of the metrics across 80 epochs. A stronger scientific validation would evaluate a held-out test set containing geographically diverse rooftops and report mean and standard deviation across images. Comparisons against standalone CNN segmentation models, modern Transformer segmentation networks, and dedicated object detectors would also clarify the incremental benefit of the hybrid architecture.

6.1 Practical Significance

Despite these limitations, the system demonstrates a useful progression from remote-sensing imagery to actionable solar information. The ability to detect PV regions, localize them, estimate their area, and calculate approximate energy potential can support rooftop inventory creation, preliminary solar-resource assessment, renewable-energy planning, and visual monitoring. The modular architecture also makes it possible to replace individual components as stronger models, higher-resolution imagery, or richer geographic information become available.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

6.2 Reproducibility and Implementation Characteristics

From Table 5, the implementation is based on widely used Python tools and explicit model parameters. The image size is 256 × 256, training uses batch size four, 80 epochs, learning rate 0.0001, Adam optimization, and BCELoss, while the Transformer uses d_model 512, eight attention heads, and two layers. The application uses a 0.50 prediction threshold and defaults to 211.73 W/m² power density, five peak-sun hours per day, and 75% system efficiency. These details make the described workflow sufficiently concrete for reconstruction of the prototype described in the thesis.

Table 5: Model and System Configuration Parameters

Parameter	Implemented setting
Input size	256 × 256 pixels
CNN backbone	Pretrained ResNet-18
Transformer	2 encoder layers
Embedding dimension	512
Attention heads	8
Batch size	4
Epochs	80
Learning rate	0.0001
Optimizer	Adam
Loss	Binary Cross-Entropy
Prediction threshold	0.50
Power density	211.73 W/m²
Peak sun hours	5 hours/day
System efficiency	75%
Default spatial resolution	0.15 m/pixel

VII. CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

This paper presented a Hybrid Transformer-Based Solar PV Panel Detection and Energy Estimation System for automated analysis of rooftop imagery. The proposed framework combines the spatial feature-learning capability of a ResNet-18 CNN backbone with the global contextual modelling capability of a Transformer Encoder. The resulting representation is decoded into a binary segmentation mask, which is subsequently converted into localized panel regions using contour extraction. This design enables the system to provide both pixel-level segmentation and object-level localization.

The work further extends detection into quantitative solar analysis. Segmented PV pixels are converted into physical panel area using spatial resolution, and the estimated area is used to calculate approximate installed capacity and daily electricity generation. In the reported example, the two detected regions cover 30.52 m² and 23.31 m², producing approximately 11.40 kW combined capacity and 42.74 kWh/day estimated generation under the adopted assumptions. The model reports IoU 0.841, Dice 0.914, precision 0.982, and recall 0.855 for the test-image example, with the corresponding confusion matrix yielding approximately 0.993pixel accuracy.

The training curves indicate progressive improvement across 80 epochs, with the later stages showing comparatively stable Dice, IoU, precision, and recall behaviour. The Streamlit application integrates the full pipeline into a single interactive environment and provides image visualization, segmentation masks, bounding boxes, coordinates, evaluation metrics, training plots, and solar-analysis outputs. Overall, the developed system demonstrates the feasibility of combining hybrid deep-learning segmentation with practical solar-resource estimation in an accessible application framework.

7.2 Future Scope

Future work can improve the system in several directions. More advanced Transformer architectures, including Vision Transformers and Swin Transformer-based segmentation models, can be evaluated to determine whether stronger contextual representations improve the detection of small or partially obscured panels. Larger and geographically diverse datasets can also improve generalization and provide a stronger basis for statistical evaluation across different rooftop styles, illumination conditions, and environments.

The localization stage can be extended from connected-region bounding boxes toward instance segmentation so that adjacent panels can be separated more accurately. Higher-resolution satellite and UAV imagery, together with GIS and building-footprint information, can improve spatial precision and enable more detailed mapping. Repeated imagery could also support change detection and monitoring of newly installed or removed PV systems.

The energy-estimation component can be made more realistic by incorporating irradiance, weather, seasonal variation, panel orientation, tilt, shading, temperature, and module or inverter characteristics. Rather than relying on fixed peak-sun-hour and efficiency values, future versions could combine detected panel geometry with site-specific solar-resource information. This would allow the system to move from a

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

potential-energy estimate toward a more calibrated generation forecast.

Finally, the Streamlit platform can be expanded with interactive geospatial maps, automated report generation, historical comparison, rooftop suitability scoring, installation planning, and large-scale batch processing. Computational optimization through pruning, quantization, or model compression may also support deployment on resource-constrained edge devices. These extensions would move the current research prototype toward a scalable platform for automated rooftop solar assessment and renewable-energy planning.

REFERENCES

[1] J. García, M. Fernandez, and L. Gomez, "Generalized deep learning model for photovoltaic module segmentation in aerial imagery," *Solar Energy*, vol. 275, pp. 112–125, 2024.

[2] Y. Zhang, H. Li, and X. Wang, "Deep learning-based rooftop photovoltaic extraction framework using aerial imagery and building footprint data," *ISPRS International Journal of Geo-Information*, vol. 14, no. 3, pp. 101–118, 2025.

[3] M. Kruitwagen, D. M. J. Tax, and R. van de Weijer, "Space-scale exploration of the poor reliability of deep learning models: The case of rooftop photovoltaic detection," *Environmental Data Science*, vol. 4, pp. 1–20, 2025.

[4] S. Khan, A. Rehman, and M. Shah, "A comprehensive review of remote sensing and artificial intelligence for image analysis," *Sensors*, vol. 25, no. 19, pp. 5965–5988, 2025.

[5] R. Sharma, P. Kumar, and S. Singh, "Survey of photovoltaic panel detection methods using image processing and deep learning," *Renewable and Sustainable Energy Reviews*, vol. 178, pp. 113–130, 2024.

[6] L. Chen, Q. Wang, and Y. Li, "Deep learning in remote sensing image classification: A review," *Neural Computing and Applications*, vol. 36, pp. 10235–10260, 2024.

[7] H. Zhao, X. Liu, and Y. Chen, "CPVPD-2024: A large-scale photovoltaic power plant dataset and deep learning detection framework," *Scientific Data*, vol. 12, pp. 145–160, 2025.

[8] A.Patel, S. Mehta, and R. Gupta, "Deep learning-based rooftop photovoltaic detection and energy assessment using aerial imagery," *Sustainability*, vol. 17, no. 15, pp. 6853–6870, 2025.

[9] T. Nguyen, J. Lee, and K. Park, "Comparative study of deep learning models for photovoltaic panel detection in aerial imagery," *IEEE Access*, vol. 13, pp. 45021–45035, 2025.

[10] F. Müller, S. Bosch, and T. Erbertseder, "Monitoring rooftop photovoltaic development using deep learning and aerial imagery," *Applied Energy*, vol. 356, pp. 121981, 2024.

[11] P. Li, X. Zhang, and Y. Wang, "Automated detection and tracking of photovoltaic modules using 3D remote sensing data," *Applied Energy*, vol. 348, pp. 120–135, 2024.

[12] M. Santos, R. Torres, and J. Silva, "Fast automatic solar module geo-labeling from UAV thermal imagery using deep learning," *Solar Energy Advances*, vol. 5, pp. 100–115, 2025.

[13] K. Patel, A. Sharma, and R. Singh, "Solar photovoltaic assessment using large language models and computer vision," *arXiv preprint* arXiv:2507.19144, 2025.

[14] D. Kim, H. Park, and J. Kim, "Enhanced rooftop photovoltaic detection using CNN feature aggregation techniques," *arXiv preprint* arXiv:2501.02840, 2025.

[15] Y. Zhou, Z. Wang, and Q. Li, "Survey of deep learning methods for remote sensing satellite image analysis," *IEEE Geoscience and Remote Sensing Review*, vol. 61, pp. 1–20, 2023.

[16] S. Gupta, A. Verma, and R. Kumar, "AI-based solar panel detection using high-resolution satellite imagery," *Journal of Renewable Energy*, vol. 2025, pp. 1–12, 2025.

[17] M. Rahman, K. Ahmed, and J. Kim, "Solar photovoltaic potential assessment using deep learning and remote sensing," *Energy Reports*, vol. 11, pp. 245–260, 2025.

[18] L. Wang, Y. Chen, and H. Li, "Deep learning-based object detection in remote sensing imagery: A review," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[19] J. Peters, M. Brown, and S. Clark, "Automated rooftop solar panel detection using convolutional neural networks," *Remote Sensing*, vol. 16, no. 4, pp. 850–865, 2024.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 36-48, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309005
Available Online at https://www.jaiet.com

Citation of this Article:

J. S. S. Prasanna, & S. Chandra Sekhar. (2026). Hybrid Transformer-Based Solar PV Panel Detection and Energy Estimation System. *Journal of Artificial Intelligence and Emerging Technologies (JAIET).* 3(9), 36-48. Article DOI: https://doi.org/10.47001/JAIET/2026.309005

***** End of the Article *****