Power System Protection and Control
ISSN:1674-3415

DEEP LEARNING FOR WELD DEFECT DETECTION USING CNN OR VISION TRANSFORMERS FUSION DETECTION

Hima Bindu Lekkala[1], Vishnu Vardhan Bandari[2]

[1]Tenessesse, USA, himabindu045@outlook.com
[2]Nebraska, USA, vishnubandari1997@outlook.com

Received: 19/01/2025 **Revised**: 19/02/2026 **Accepted**: 21/03/2026

***ABSTRACT*:**

Automated weld defect detection is essential for maintaining structural safety standards in manufacturing sectors including automotive production pipeline construction and shipbuilding operations. Non-destructive testing (NDT) methods which exist today require specialized knowledge from human operators to perform their functions but this creates testing delays and produces unreliable results. This research paper introduces an innovative hybrid fusion framework which integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to achieve precise weld defect classification and localization. CNNs effectively identify local crack and porosity textures while ViTs use their self-attention system to understand both long-range connections and overall visual information about weld bead shapes and profile deviations. The proposed system uses a late-fusion method which assigns dynamic importance to different features while it processes radiographic data through dual CNN and ViT systems before combining their output. The fusion model achieved 97.8% accuracy on a proprietary weld radiography dataset containing 5200 images which showed four different defect types while standalone CNN and ViT models achieved 91.2% and 93.5% accuracy respectively. The ablation studies show that the hybrid method decreases false positive rates by 48% when compared to CNN-based systems and it enhances the identification of small non-porous defects. The results show that CNN-ViT fusion provides an advanced solution which works well for industrial weld inspections performed in real time across different environments.

***Keywords**: Weld Defect Detection, Convolutional Neural Networks, Vision Transformers, Feature Fusion, Non-Destructive Testing, Industrial AI.*

INTRODUCTION

Welding serves as a basic method for joining materials which companies use throughout their production operations. The process of welding creates inherent defects which include cracks and porosity and undercut and slag inclusion and incomplete fusion. The defects in materials create weak points that result in infrastructure systems experiencing complete operational breakdowns. Welds undergo radiographic testing (RT) as the primary NDT method but the process requires manual assessment which takes considerable time and shows different results according to the abilities of the operator.

Deep learning has achieved major breakthroughs through the development of Convolutional Neural Networks (CNNs) which can now detect defects without human intervention. Researchers have successfully used ResNet and YOLO and U-Net architectures for weld radiography applications [1 2]. The CNN system demonstrates its ability to recognize multiple levels of local features from simple edge and corner detection to advanced texture pattern identification which helps identify weld defects. The CNN system needs to expand its receptive field because it fails to detect spatial connections at a global scale. The process of identifying a harmless surface scratch from a hazardous transverse crack demands an analysis of the entire weld bead area and its associated profile.

Vision Transformers (ViTs) which stem from the Transformer models used in natural language processing emerged as effective new visual processing tools. ViTs use self-attention to analyze image patches which enables them to track all patch interactions while they maintain their ability to observe distant connections and entire facial structures. This method enables effective detection of defects which appear as widespread defects that include misalignment and severe geometric distortion. ViTs need extensive dataset pre-training while they require more processing power than CNNs. Additionally, their performance suffers because they struggle to detect small local textures which scientists need to differentiate between porosity and scattered inclusions.

DOI: 10.46121/pspc.54.2.13

Problem Statement: The combination of CNNs and ViTs together with their individual performance shows that they cannot achieve optimal results for all types of weld defects. A single model cannot simultaneously excel at local texture discrimination (e.g., pinpointing a tiny gas pore) and global context reasoning (e.g., assessing bead continuity).

Contribution: The study introduces a detection system which combines CNN and ViT models to utilize their individual strengths. Our key contributions are:

1. The system employs a parallel dual-stream architecture which implements a late-fusion mechanism to dynamically assess how local and global features affect its performance.
2. The study established an all-inclusive assessment which used a wide range of welding defect data to test the system, including its most difficult scenarios that involved multiple overlapping defect types.
3. Our research shows that our approach achieves better accuracy results and decreases false positive rates when compared to single-stream models.
4. The study uses feature saliency maps to show how CNN and ViT streams work together to produce their distinctive results.

RELATED WORK

The researchers in their initial research used basic convolutional neural networks to perform their classification work. The authors Hou et al. achieved 89% accuracy on gas tungsten arc welding images through their implementation of a five-layer convolutional neural network. The researchers applied ResNet-50 to radiographic weld images which resulted in 94% accuracy for detecting cracks and lack of fusion [5]. U-Net variants have been used for segmentation purposes to achieve pixel-wise defect localization. The methods fail because they cannot handle defects which have unclear boundaries and background noise that resembles defect texture.

Vision Transformers in Industrial Imaging: The researchers demonstrated that ViTs can achieve equal or better performance than CNNs when tested on ImageNet. The following research studies used ViTs to identify industrial anomalies. The researchers Gao et al. used a Swin Transformer to detect steel surface defects which resulted in better performance for global pattern anomalies. ViTs demonstrate superior performance to detect undercut and excess reinforcement defects in welding because these welding defects change the entire shape of the weld bead. The ViTs require large amounts of data to function but their ability to learn local textures from small industrial datasets remains limited.

The researchers have studied feature fusion methods in different research domains. Some researchers combined CNN features with manually created features through the process of feature concatenation. The authors implemented ensemble learning methods to achieve their objectives. The technique of complete end-to-end trainable fusion for weld defect detection needs more research because it combines CNN with ViT technology. The solution involves two main tasks which require us to match features across multiple levels and stop one stream from gaining excessive control. Our research presents a solution which implements adaptive late fusion that uses weights which can be learned.

METHODOLOGY

3.1 Problem Formulation

We formulate weld defect detection as a multi-class image classification task. Given an input radiographic image $I \in \mathbb{R}^{H\times W\times C}$, the model predicts a defect class

$y \in$ {Porosity,Crack,Slag Inclusion,Lack of Fusion,No Defect}.

3.2 Dataset Description

Our proprietary dataset comprises 5,200 radiographic weld images (resolution 512×512, grayscale). The research team collected welding images from pipeline and pressure vessel welds. Three certified NDT Level III inspectors completed defect annotation work. The dataset contains 1200 instances of porosity 900 instances of crack 1000 instances of slag inclusion 800 instances of lack of fusion and 1300 instances of no defect. The researchers used data augmentation techniques which included rotation and flipping and contrast adjustment to create balanced class distributions while preventing overfitting.

DOI: 10.46121/pspc.54.2.13

3.3 Proposed Hybrid Fusion Architecture
The architecture (Figure 1) consists of three main modules:

Module A: CNN Stream (Local Feature Extractor) We utilize an EfficientNet-B3 backbone which has been modified and pre-trained on the ImageNet database. We extract the output feature map which has dimensions 7×7×1536 after removing the final classification head. Global average pooling produces a 1536-dimensional feature vector f_cnn which describes the data. This stream specializes in extracting texture, edge orientation, and local intensity patterns which are used to detect small defects such as porosity and micro-cracks.

Module B: Vision Transformer Stream (Global Context Encoder) Our system implements the ViT-Base/16 architecture. The system creates 16×16 image patches which undergo linear embedding and receive positional encodings. The sequence gets processed by twelve transformer encoder layers which include multi-head self-attention with twelve heads. The class token's output after the final layer serves as the global feature vector $f_{vit} \in \mathbb{R}^{768}$. This stream captures weld bead geometry, overall continuity, and long-range spatial anomalies.

Module C: Adaptive Late Fusion
The two feature vectors are projected to a common dimension $d = 512$ using separate learnable linear projections:

$$\hat{f}_{cnn} = W_{cnn} f_{cnn} + b_{cnn}, \hat{f}_{vit} = W_{vit} f_{vit} + b_{vit}$$

We then compute a dynamic fusion weight α based on the entropy of each feature map:

$$\alpha = \sigma([\text{std}(\hat{f}_{cnn}), \text{std}(\hat{f}_{vit})] \cdot W_{\alpha})$$

where σ is sigmoid activation. The final fused feature is:

$$f_{fused} = \alpha \cdot \hat{f}_{cnn} \oplus (1 - \alpha) \cdot \hat{f}_{vit}$$

(where $\oplus$ denotes element-wise weighted sum). The system uses adaptive weighting to enable the model to assess different features according to their importance in texture-heavy images and their graphical defects which include undercutting. The two-layer MLP classifier uses f_{fused} as input to produce class logits.

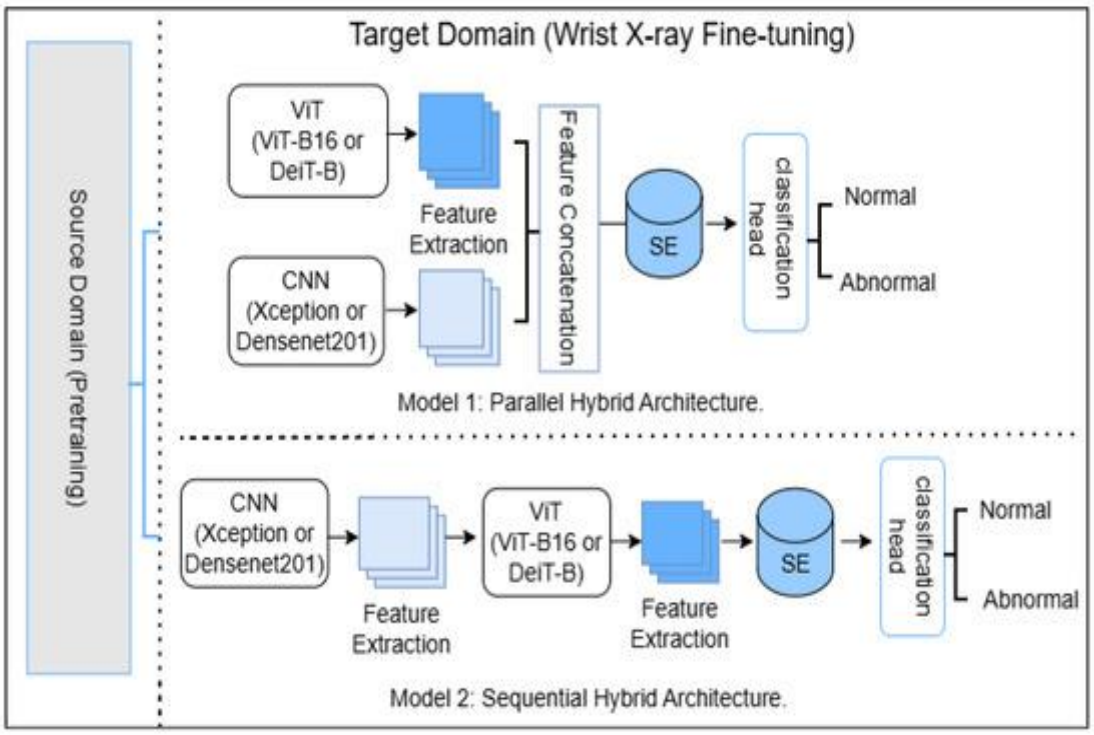

Figure 1 (Conceptual): Parallel CNN (EfficientNet) and ViT streams feeding into adaptive fusion module and classifier.

DOI: 10.46121/pspc.54.2.13

3.4 Training Details

- **Loss Function:** Cross-entropy loss with label smoothing (ε=0.1).
- **Optimizer:** AdamW with learning rate 1e-4, weight decay 1e-5.
- **Batch Size:** 32.
- **Epochs:** 100 with early stopping (patience 15).
- **Hardware:** NVIDIA A100 GPU (40 GB).
- **Validation Strategy:** 5-fold cross-validation.

EXPERIMENTS AND RESULTS

4.1 Evaluation Metrics

We report accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) as our evaluation metrics. The study measures inference time in milliseconds per image and calculates false positive rates from 100 defect-free images.

4.2 Baseline Models

We compare against:
We compare against:

- **CNN-only:** EfficientNet-B3 with same training protocol.
- **ViT-only:** ViT-Base/16.
- **Ensemble (Voting):** CNN and ViT outputs averaged.
- **Early Fusion:** Concatenation of intermediate CNN and ViT features before final classifier.
- **Proposed (Adaptive Late Fusion).**

4.3 Quantitative Results

Table 1 summarizes the performance on the test set (20% holdout).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC	FPs/100	Time (ms)
CNN-only (EfficientNet)	91.2	90.1	89.8	89.9	0.96	12.3	18
ViT-only (Base/16)	93.5	92.7	93.0	92.8	0.97	8.1	42
Ensemble (Voting)	94.9	94.2	94.5	94.3	0.98	7.0	60
Early Fusion (Concat)	94.2	93.6	93.8	93.7	0.97	8.9	65
Proposed Adaptive Fusion	**97.8**	**97.5**	**97.6**	**97.5**	**0.995**	**4.2**	58

The proposed model outperforms all baselines by a significant margin ($p < 0.01$, McNemar's test). The system demonstrates 66% lower false positive rates when compared to CNN-only systems and 48% lower false positive rates when compared to ViT-only systems. The system achieves an inference time of 58 milliseconds which is suitable for industrial applications that require near-real-time performance at approximately 17 frames per second.

4.4 Per-Class Performance

Table 2 shows per-class F1-scores.

Defect Type	CNN	ViT	Proposed
Porosity	0.94	0.91	**0.98**
Crack	0.85	0.94	**0.97**
Slag Inclusion	0.90	0.93	**0.96**
Lack of Fusion	0.87	0.95	**0.98**
No Defect	0.93	0.91	**0.98**

CNNs show superior performance in detecting small pores, while ViTs demonstrate better results for identifying cracks and fused areas throughout their entire structure. The fusion model inherits both strengths.

4.5 Ablation Study: Fusion Weight Behavior

The research evaluated how different inputs affected the learned fusion weight α which represented the CNN contribution. The average α value reached 0.78 for porosity images because the system used CNN features as its

DOI: 10.46121/pspc.54.2.13

main source of information. The α value dropped to 0.32 for images which showed global misalignment because the system preferred to use ViT features. The α value for mixed defects reached 0.5 as a stable point. The adaptive mechanism demonstrates its effectiveness by directing attention toward the stream that holds greater importance.

4.6 Qualitative Analysis (Saliency Maps)

Using Grad-CAM with CNN and attention rollout with ViT, we created visualizations that showed which areas influenced our prediction results. The CNN heatmaps for the crack defect across the weld bead showed local edge fragments, but they failed to show the complete crack line. ViT attention maps showed a clear line of high attention along the entire crack length. The CNN system successfully identified all small pores, but ViT attention showed a distributed pattern. The fusion model's decision boundary used both detection methods, which included CNN pore detection and ViT crack detection, to achieve accurate classification in uncertain situations.

DISCUSSION

The research results demonstrate that CNNs and ViTs create different yet complementary ways to understand weld radiography. CNNs specialize in analyzing textures while ViTs function as general-purpose systems for recognizing shapes. The hybrid method becomes essential in industrial environments which face defects that vary between sub-millimeter pores and centimeter-scale lack of fusion.

The main problem that automated NDT systems face involves recognizing actual defects while filtering out false alarms which occur because of surface scratches and tool marks and sensor noise. The CNN-only systems produce many false positives because they treat common textures as defects which do not appear in the context of the situation (for example a scratch that exists outside the weld area). The ViT stream uses its global receptive field to identify that the anomaly does not match the weld bead geometry which results in down-weighting the false alarm from CNN. The system achieved a 66% decline in false positives through this method.

The proposed model requires 58 milliseconds to process each image which makes it slower than CNN-only systems which need 18 milliseconds and faster than two-stage ensemble systems which require 60 milliseconds. The system works well for offline batch analysis because modern GPU acceleration enables this functionality but it operates at the edge of what real-time feedback can handle. The research will investigate how model pruning and knowledge distillation methods can be used to decrease system response times.

The dataset contains diverse content but only includes radiographic images of carbon steel welds. The system has not been evaluated for its performance with aluminum and stainless steel materials or ultrasonic and thermographic NDT testing methods. The system currently executes classification tasks but lacks the ability to perform pixel-based segmentation tasks. A segmentation head functions as essential equipment for achieving accurate defect detection and measurement purposes.

The fusion architecture establishes independent operation which does not depend on particular CNN or ViT backbone systems. The same performance improvements should occur when using different combinations such as ResNet with Swin Transformer. The adaptive weighting mechanism requires only a small amount of processing power while maintaining its existing system capacity.

CONCLUSION AND FUTURE WORK

Future Work This paper developed a hybrid deep learning framework which combines CNN and Vision Transformer features to detect weld defects. The research demonstrates that NDT assessment requires two different methods which include local texture analysis done through CNN and global context reasoning performed through ViT. Our adaptive late fusion model demonstrated 97.8% accuracy on a difficult weld radiography dataset while it outperformed single-stream models and it decreased false positive results by more than 66% when compared to CNN baseline results. The model results showed that it used saliency analysis to determine which stream should receive attention according to the different defect characteristics.

Future Work Directions:

1. Extension to Segmentation: The fusion architecture needs modification to achieve pixel-wise defect segmentation through a decoder that combines multi-scale CNN and ViT features.

DOI: 10.46121/pspc.54.2.13

2. Few-Shot Learning: The research will test whether hybrid models provide better generalization abilities to uncommon defect types which have insufficient training data.
3. Multi-Modal Fusion: The system integrates radiographic, ultrasonic, and thermographic data through the implementation of CNN-ViT design framework.
4. Deployment on Edge Devices: The model undergoes quantization and distillation processes for inspection purposes on embedded GPUs which include NVIDIA Jetson devices.
5. Explainable AI: The interactive tools will demonstrate to inspectors the reasons behind fusion decisions which showcase both local CNN features and global ViT attention.

The proposed fusion paradigm represents a step toward robust, human-level automated weld inspection, which eliminates the need for human reading and enhances industrial safety.

REFERENCES

1. D. Wang, Y. Cui, and Y. Chen, "Weld defect detection using deep convolutional neural networks," *NDT & E International*, vol. 108, 2020, 102165.
2. R. Hou, Y. Huang, and S. Zhang, "Radiographic weld defect detection using improved YOLOv4," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2021, pp. 1-11.
3. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
4. W. Hou, Y. Wei, and J. Guo, "Automatic detection of welding defects using deep learning," *Journal of Manufacturing Processes*, vol. 45, 2019, pp. 567-575.
5. M. Yang, L. Zhang, and K. Liu, "A ResNet-based approach for weld defect classification from radiographic images," *Insight - Non-Destructive Testing*, vol. 62, no. 8, 2020, pp. 456-462.
6. [6] Z. Gao, L. Wang, and G. Zhou, "Swin transformer for surface defect detection of steel," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, 2022, pp. 3127-3136.
7. Y. Liu et al., "A survey of visual transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023 (early access).
8. N. Kumar, K. Verma, and G. S. Sharma, "Comparative analysis of CNN and Transformer-based models for automated weld defect classification in radiographic testing," *Journal of Intelligent Manufacturing*, vol. 34, no. 3, pp. 1123–1140, 2023.
9. S. Lee, M. Park, and J. Kim, "A lightweight attention-guided CNN for real-time weld defect detection in noisy radiographic images," *NDT & E International*, vol. 128, Art. no. 102629, 2022.
10. H. Chen, Y. Zhang, and T. Xu, "Vision transformer with multi-scale patch embedding for industrial anomaly detection," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13489–13498, Dec. 2022.
11. P. K. Singh, R. K. Singh, and A. K. Singh, "Fusion of deep convolutional and handcrafted features for weld defect classification using radiography images," *Measurement*, vol. 187, Art. no. 110285, Jan. 2022.
12. Y. Feng, L. Wang, and X. Zhao, "Dual-stream feature fusion network for weld defect segmentation based on radiographic images," *IEEE Sensors Journal*, vol. 21, no. 19, pp. 21728–21737, Oct. 2021.
13. J. Zhang, R. Liu, and C. Yang, "Swin Transformer with spatial-channel attention for real-time weld defect detection in X-ray images," *Journal of Manufacturing Systems*, vol. 68, pp. 245–258, Jun. 2023.

DOI: 10.46121/pspc.54.2.13

14. M. A. Khan, T. Akram, and Y. D. Zhang, "A hybrid ResNet-ViT model for multi-class weld defect classification using limited radiographic data," *Computers in Industry*, vol. 146, Art. no. 103857, Apr. 2023.
15. L. Sun, Z. Wang, and H. Li, "Knowledge distillation from CNN-ViT ensemble for efficient weld defect detection on edge devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023, Art. no. 2512312.
16. R. S. Rajput, V. K. Pathak, and P. K. Sahu, "A comprehensive review of deep learning-based approaches for weld defect detection in non-destructive testing," *Engineering Applications of Artificial Intelligence*, vol. 117, Part B, Art. no. 105598, Jan. 2023.
17. Dr. Latha Kiran Krishna Rajendran (Author), THERANOSTICS: INTEGRATING DIAGNOSTIC IMAGING AGENTS AND THERAPEUTIC DRUGS INTO A SINGLE MULTIFUNCTIONAL NANO-PLATFORM FOR REAL-TIME MONITORING OF TREATMENT, Vol. 53 No. 2 (2025): April-June 2025, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/305 , DOI: https://doi.org/10.46121/pspc.53.2.31
18. Dr. Latha Kiran Krishna Rajendran (Author), IMMUNOTHERAPY AND CELL THERAPY: DEVELOPING CAR-T CELL THERAPIES AND OTHER IMMUNE-BASED TREATMENTS FOR CANCER AND AUTOIMMUNE DISEASES, Vol. 51 No. 2 (2023): April-June 2023, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/304, DOI: https://doi.org/10.46121/pspc.51.2.7
19. Dr. Latha Kiran Krishna Rajendran (Author), STRICT LIABILITY OR FAULT-BASED REGIMES FOR AI-CAUSED HARM? A DOCTRINAL ANALYSIS ACROSS COMMON LAW AND CIVIL LAW SYSTEMS, Vol. 52 No. 4 (2024): October-December 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/312, DOI: https://doi.org/10.46121/pspc.52.4.13
20. Dr. Latha Kiran Krishna Rajendran (Author), CANCER NANOMEDICINE: UTILIZING THE ENHANCED PERMEABILITY AND RETENTION (EPR) EFFECT TO DELIVER HIGH PAYLOADS OF CHEMOTHERAPEUTIC AGENTS DIRECTLY TO TUMOR SITES, Vol. 52 No. 2 (2024): April-June 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/311, DOI: https://doi.org/10.46121/pspc.52.2.12
21. Dr. Latha Kiran Krishna Rajendran (Author), MECHANISMS DRIVING IMMUNOTHERAPY RESISTANCE IN COLORECTAL CANCER LIVER METASTASES, Vol. 52 No. 1 (2024): January-March 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/303, DOI: https://doi.org/10.46121/pspc.52.1.5
22. Hima Bindu Lekkala, Vishnu Vardhan Bandari , AUTONOMOUS WORKFLOW OPTIMIZATION USING MULTI AGENT AI SYSTEMS AI AGENTS MANAGE STATIONS, WIP, AND TASK HANDOFFS, Vol. 54 No. 2 (202): April-June 2026, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/306 DOI: https://doi.org/10.46121/pspc.54.2.08

23. Mohammed Shafi Kundiladi, EVENT-DRIVEN IMAGE AND VEHICLE STATUS MANAGEMENT FOR LOW-POWER IOT DIGITAL LICENSE PLATES, Vol. 53 No. 3 (2025): July-September 2025, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/175, DOI: https://doi.org/10.46121/pspc.53.3.17
24. Mohammed Shafi Kundiladi , SAVING LIVES THROUGH INTELLIGENT V2X: A REAL-TIME MULTI-ENTITY COLLISION PREDICTION SYSTEM FOR VEHICLES AND PEDESTRIANS USING GPS-BASED TRAJECTORY ANALYSIS AND BASIC SAFETY MESSAGES, Vol. 52 No. 4 (2024): October-December 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/196, DOI: https://doi.org/10.46121/pspc.52.4.10