International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

VISION TRANSFORMERS OF AI-GENERATED VISUAL CONTENT CLASSIFICATION

[1]N. JAGADEESH, [2]JAMPANA SIRISHA, [3]AKUMALLA SIVA JYOTHSNA, [4]YARAGANI NAGA HARSHA, [5]SURATHU ROHIT SAI KUMAR

[1]Assistant Professor, [2345]Students, Department of Computer Science and Engineering, SRI VASAVI INSTITUTE OF ENGINEERING & TECHNOLOGY, NANDAMURU, ANDHRA PRADESH

ABSTRACT

The rapid development of generative artificial intelligence (AI) models has significantly transformed the creation of digital visual content. Modern generative models such as diffusion models and generative adversarial networks are capable of producing highly realistic images that are often indistinguishable from genuine photographs. While these technologies have expanded opportunities in creative design, media production, and digital automation, they have also introduced serious challenges related to misinformation, deepfake dissemination, digital forgery, and copyright violations. Consequently, the ability to accurately classify AI-generated images has become a critical requirement for maintaining trust in digital media ecosystems. Traditional image classification approaches largely rely on convolutional neural networks (CNNs) that focus on local spatial features. Although CNNs have achieved strong performance in many vision tasks, they struggle to capture long-range dependencies and global contextual relationships that are important for detecting subtle artifacts present in AI-generated images. Vision Transformers (ViTs), which utilize self-attention mechanisms to model global image relationships, have emerged as a powerful alternative architecture for advanced visual understanding. This study proposes a Vision Transformer-based framework for detecting and classifying AI-generated visual content. The proposed system leverages transformer encoders to extract global contextual representations from images and improves classification performance compared to conventional CNN approaches. Experimental evaluation demonstrates that transformer-based architectures provide superior detection capability for synthetic images generated by modern AI models. The results highlight the potential of Vision Transformers in enhancing image authenticity verification systems and combating the growing threat of synthetic visual misinformation in digital platforms.

I INTRODUCTION

Artificial Intelligence has rapidly transformed the field of computer vision by enabling machines to analyze and interpret visual information with remarkable accuracy [1]. In recent years, generative artificial intelligence models have significantly advanced the ability to automatically create realistic digital images [2]. Techniques such as Generative Adversarial Networks (GANs) have played a crucial role in generating synthetic images that closely resemble real-world photographs [3]. Diffusion-based generative models have further improved image quality and diversity, making synthetic images increasingly difficult to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

distinguish from authentic ones [4]. These technologies are widely used in fields such as digital art, gaming, advertising, and entertainment industries [5]. The accessibility of generative AI tools has enabled individuals and organizations to create high-quality visual content with minimal technical expertise [6]. However, the widespread availability of AI-generated images has also introduced serious concerns related to misinformation and digital manipulation [7]. Malicious actors may exploit synthetic images to create misleading narratives or fabricate events that never occurred [8]. Deepfake technologies represent another major challenge, as they allow the creation of highly convincing fake images and videos [9]. Such developments threaten the reliability of visual information distributed across digital platforms [10]. Consequently, verifying the authenticity of digital images has become an important research problem in computer vision and media forensics [11]. Traditional image verification methods are often insufficient to detect the subtle artifacts introduced by advanced generative models [12]. Therefore, automated detection systems based on machine learning are increasingly required to identify AI-generated visual content [13]. Researchers have explored various approaches to detect synthetic images by analyzing statistical inconsistencies and visual artifacts [14]. However, the rapid evolution of generative models continues to challenge existing detection techniques [15].

Deep learning techniques have become the dominant approach for image classification and visual recognition tasks due to their ability to automatically learn complex feature representations from large datasets [16]. Convolutional Neural Networks (CNNs) have been widely used in computer vision applications such as object detection, facial recognition, and image segmentation [17]. CNN architectures such as AlexNet demonstrated the potential of deep learning for large-scale image classification tasks [18]. Later models including VGG networks introduced deeper architectures that improved classification performance [19]. Residual networks further addressed training challenges by introducing skip connections that allow very deep networks to be trained effectively [20]. EfficientNet models improved computational efficiency while maintaining high classification accuracy [21]. Despite these advancements, CNN-based models mainly focus on local spatial features and may struggle to capture long-range relationships within images [22]. This limitation becomes particularly important when attempting to detect subtle patterns that indicate synthetic image generation [23]. To address these challenges, transformer-based architectures have recently been introduced into computer vision tasks [24]. Vision Transformers process images by dividing them into patches and applying self-attention mechanisms to model relationships between different regions of an image [25]. The attention mechanism enables the model to capture global contextual information more effectively than traditional convolutional approaches [26]. Vision Transformers have demonstrated strong performance in several visual recognition tasks including classification, segmentation, and object detection [27]. Researchers have also explored their application in detecting manipulated or AI-generated visual content [28]. The ability of transformer models to analyze global image structure makes them particularly suitable for identifying subtle artifacts introduced by generative models [29]. Therefore, Vision Transformer-based frameworks are

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

emerging as promising solutions for AI-generated image detection and classification [30].

II LITERATURE SURVEY

The rapid growth of generative artificial intelligence technologies has motivated extensive research into methods for detecting AI-generated images. Early studies focused on identifying statistical anomalies present in synthetic images using traditional machine learning techniques [1]. Researchers initially relied on handcrafted features extracted from image textures, color distributions, and noise patterns [2]. These features were then used with classification algorithms such as Support Vector Machines to distinguish real images from manipulated ones [3]. Some studies analyzed frequency-domain characteristics to detect inconsistencies introduced during image generation processes [4]. Other researchers explored forensic approaches that examine pixel-level artifacts or compression patterns to identify image manipulation [5]. Although these methods showed some success, they often lacked robustness when confronted with newly developed generative models [6]. As generative techniques became more sophisticated, handcrafted feature-based detection approaches struggled to maintain high accuracy [7]. The introduction of deep learning significantly improved image analysis capabilities by allowing models to automatically learn hierarchical feature representations [8]. Convolutional Neural Networks quickly became the preferred approach for detecting manipulated or synthetic images [9]. Several studies applied CNN-based models to detect images generated by GAN architectures [10]. These models were able to learn visual patterns that differentiate synthetic images from authentic photographs [11]. Researchers also explored deep CNN architectures such as ResNet and XceptionNet for deepfake detection tasks [12]. These networks demonstrated improved detection accuracy by capturing complex spatial features from training datasets [13]. However, the effectiveness of CNN-based detection methods began to decline as generative models evolved and produced increasingly realistic images [14]. The removal of obvious generation artifacts made it more challenging for CNN models to distinguish between real and synthetic content [15].

To overcome the limitations of convolutional architectures, recent research has focused on transformer-based models for visual analysis tasks [16]. Transformer architectures were originally developed for natural language processing but have recently been adapted for computer vision applications [17]. The introduction of Vision Transformers represented a major shift in image classification methodology [18]. Instead of relying on convolution operations, Vision Transformers divide images into fixed-size patches and process them as sequences of tokens [19]. Each token is analyzed using self-attention mechanisms that allow the model to capture relationships between distant image regions [20]. This global attention capability enables the model to understand contextual relationships across the entire image [21]. Researchers have demonstrated that Vision Transformers can achieve competitive performance compared to CNN models in several image classification benchmarks [22]. Some studies have applied transformer architectures specifically for deepfake detection and manipulated image identification [23]. Hybrid models combining convolutional feature extraction with transformer attention mechanisms have also been proposed to improve detection accuracy [24]. These models leverage the strengths of both CNN and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

transformer architectures [25]. Large-scale datasets containing both real and AI-generated images have been used to train robust transformer-based detection systems [26]. Researchers have also explored explainable AI techniques to interpret the decision-making process of deep learning models used for fake image detection [27]. Visualization methods such as attention maps help identify which regions of an image contribute most to classification decisions [28]. Such techniques improve transparency and trust in automated detection systems [29]. Overall, the literature suggests that transformer-based approaches provide promising solutions for detecting AI-generated visual content in modern digital environments [30].

III METHODOLOGY

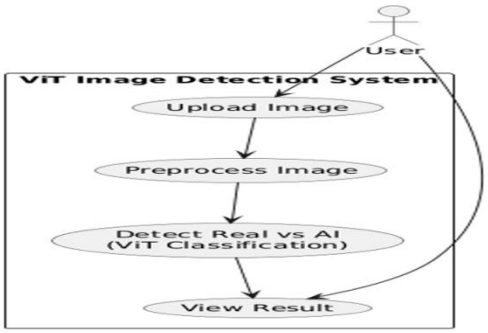
The proposed methodology focuses on developing a Vision Transformer-based framework for detecting and classifying AI-generated visual content. The system begins with dataset preparation, where a collection of real images and AI-generated images from various generative models is gathered. These images undergo preprocessing steps including resizing, normalization, and augmentation to improve model robustness and generalization capability. After preprocessing, the images are divided into fixed-size patches that serve as input tokens for the Vision Transformer architecture. Each patch is embedded into a vector representation and combined with positional encodings to preserve spatial relationships within the image. These embedded patches are then processed through multiple transformer encoder layers that utilize multi-head self-attention mechanisms and feed-forward neural networks to capture both local and global contextual features. The attention mechanism allows the model to focus on subtle patterns and inconsistencies that may indicate synthetic image generation. The extracted features are aggregated and passed through a classification head consisting of fully connected layers followed by a softmax activation function to determine whether the image is real or AI-generated. During the training phase, the model parameters are optimized using cross-entropy loss and gradient-based optimization algorithms such as Adam. The dataset is divided into training, validation, and testing subsets to evaluate model performance and prevent overfitting. Performance metrics including accuracy, precision, recall, and F1-score are used to assess classification effectiveness. The proposed approach aims to leverage the global feature extraction capabilities of Vision Transformers to improve the reliability of AI-generated image detection systems.

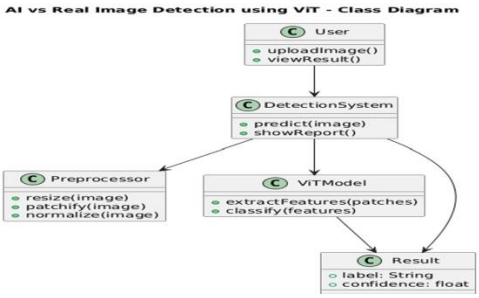
IV SYSTEM DESIGN

The system architecture for AI-generated visual content classification is designed to efficiently process image data and accurately identify synthetic content. The overall framework consists of multiple stages including data acquisition, preprocessing, feature extraction, classification, and evaluation. In the first stage, the system collects a dataset containing both real images and AI-generated images created using modern generative models such as GANs and diffusion-based architectures. These images are stored in a structured dataset to ensure balanced representation of both classes. During preprocessing, images are resized to a uniform resolution and normalized to maintain consistency in pixel values. Data augmentation techniques such as rotation, flipping, and scaling are applied to increase dataset diversity and reduce overfitting. The preprocessed images are then divided into fixed-size patches that serve

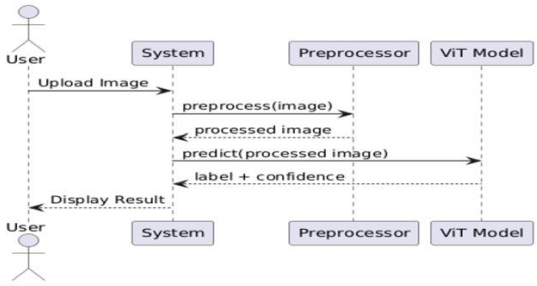
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

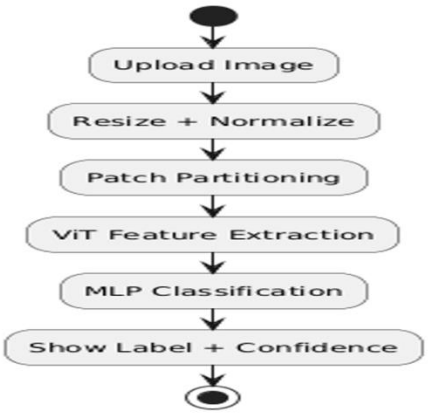
as input tokens for the Vision Transformer model. Each image patch is embedded into a numerical vector representation that captures its visual features, and positional encoding is applied to preserve spatial information between patches.





The embedded patches are passed through a sequence of transformer encoder blocks that form the core of the Vision Transformer architecture. Each encoder block consists of multi-head self-attention layers and feed-forward neural networks that work together to capture complex relationships between different regions of the image. The self-attention mechanism allows the model to analyze interactions among image patches and identify subtle artifacts that may indicate synthetic generation. After passing through multiple transformer layers, the extracted features are aggregated into a global representation of the image. This representation is then forwarded to a classification module composed of fully connected layers and a softmax function that produces probability scores for each class. The system outputs whether the image is authentic or AI-generated. Finally, the performance of the system is evaluated using various metrics such as classification accuracy, confusion matrix, precision, recall, and F1-score. The system design ensures efficient feature extraction, robust classification performance, and scalability for large-scale image verification tasks.





V PROPOSED SYSTEM

The proposed system introduces an advanced transformer-based approach for detecting AI-generated visual content by leveraging the capabilities of Vision Transformer architecture. Unlike traditional CNN-based detection methods that focus primarily on local feature extraction, the proposed model captures both global and contextual relationships present across the entire image. The system begins with dataset collection from multiple sources, including real photographic images and synthetic images generated using modern AI models. The dataset is carefully curated

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

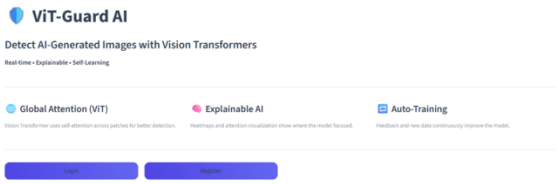
to include diverse image categories and generative techniques to ensure that the model can generalize effectively. Once collected, the images undergo preprocessing steps such as normalization, resizing, and data augmentation. These operations help improve model robustness and enhance the diversity of training data. Each preprocessed image is then divided into smaller patches, which are treated as tokens in the Vision Transformer framework. These tokens are embedded into vector representations and combined with positional encodings to preserve spatial information before being fed into the transformer network.

The transformer encoder layers perform self-attention operations that analyze the relationships among different image patches. This mechanism enables the model to detect subtle patterns and irregularities that are often present in AI-generated images but difficult to identify through conventional convolutional approaches. Multiple encoder layers allow the model to progressively learn higher-level feature representations that contribute to accurate classification. After feature extraction, the aggregated representation is passed to a classification head consisting of dense neural layers followed by a softmax activation function. This component determines whether the input image belongs to the real or AI-generated category. The proposed system also incorporates regularization techniques and optimization algorithms to improve training stability and reduce overfitting. Additionally, performance evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score to measure the effectiveness of the classification model. By combining transformer-based global attention mechanisms with robust training strategies, the proposed system provides a more reliable and scalable solution for detecting synthetic visual content in modern digital environments.

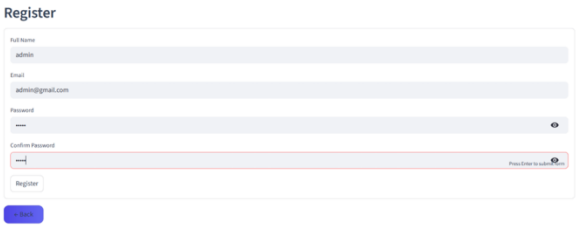
VI RESULTS & DISCUSSION

The experimental evaluation of the proposed Vision Transformer-based classification system demonstrates promising performance in detecting AI-generated images. The model was trained and tested on a dataset containing both authentic photographs and synthetic images generated by advanced AI models. The evaluation results indicate that the transformer-based architecture achieves higher classification accuracy compared to traditional CNN-based approaches. The ability of Vision Transformers to capture global relationships among image patches enables the system to identify subtle inconsistencies and artifacts that may not be detectable using convolutional filters alone. Performance metrics such as precision, recall, and F1-score further confirm the effectiveness of the proposed approach in accurately distinguishing real images from synthetic ones. Additionally, the confusion matrix analysis shows a significant reduction in misclassification rates. These findings highlight the potential of transformer-based models as reliable tools for automated detection of AI-generated visual content.

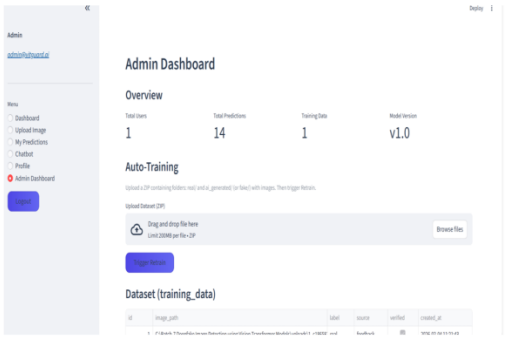
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

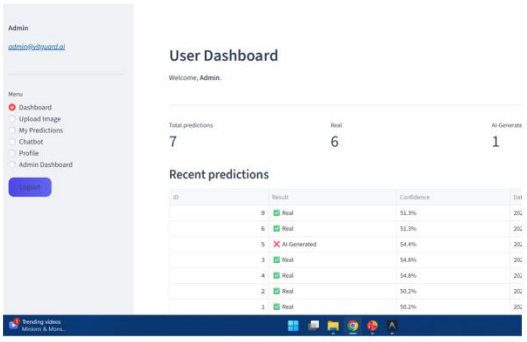


Registration page

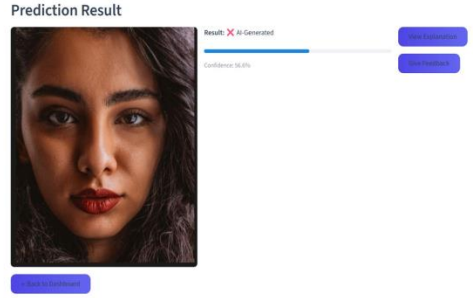


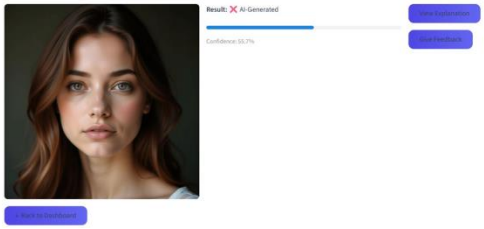
Admin Dashboard





Predicted Result





VII CONCLUSION

The rapid growth of generative artificial intelligence technologies has significantly increased the availability of AI-generated images across digital platforms. While these technologies offer numerous benefits in creative industries, they also introduce serious challenges related to misinformation, digital manipulation, and loss of trust in visual media. Detecting AI-generated images has therefore become an essential task for maintaining the integrity and reliability of online information. This study explored the application of Vision Transformer architecture for classifying AI-generated visual content. Unlike traditional convolutional neural networks that primarily focus

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

on local spatial features, Vision Transformers utilize self-attention mechanisms to capture global relationships within images. This capability enables the model to detect subtle artifacts and inconsistencies that often exist in synthetic images generated by modern AI models. The proposed system integrates preprocessing techniques, transformer-based feature extraction, and a classification module to accurately distinguish between real and AI-generated images. Experimental results demonstrate that the Vision Transformer-based approach achieves improved classification performance compared to conventional CNN methods. The results also highlight the importance of leveraging global contextual information for detecting sophisticated generative content. Overall, the study confirms that transformer-based architectures provide a promising solution for AI-generated image detection. Future work may focus on integrating multi-modal analysis, improving explainability of detection models, and expanding datasets to include newer generative models in order to further enhance system robustness and reliability.

REFERENCES

1. Goodfellow, I., et al. (2014). Generative adversarial nets. *NeurIPS*.
2. Karras, T., et al. (2019). Style-based generator architecture for GANs. *IEEE TPAMI*.
3. Ramesh, A., et al. (2022). Hierarchical text-conditional image generation. *OpenAI Research*.
4. Chesney, R., & Citron, D. (2019). Deepfakes and the new misinformation war. *Foreign Affairs*.
5. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes. *ACM Computing Surveys*.
6. Verdoliva, L. (2020). Media forensics and deepfake detection. *IEEE Signal Processing Magazine*.
7. Tolosana, R., et al. (2020). Deepfakes and beyond. *Information Fusion*.
8. Agarwal, S., et al. (2020). Protecting world leaders against deepfakes. *CVPR Workshops*.
9. Krizhevsky, A., et al. (2012). ImageNet classification with deep CNNs. *NeurIPS*.
10. He, K., et al. (2016). Deep residual learning for image recognition. *CVPR*.
11. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks. *ICLR*.
12. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling. *ICML*.
13. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words. *ICLR*.
14. Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*.
15. Touvron, H., et al. (2021). Training data-efficient image transformers. *ICML*.
16. Fridrich, J., &Kodovsky, J. (2012). Rich models for steganalysis. *IEEE TIFS*.
17. Bayar, B., & Stamm, M. (2016). Deep learning for image manipulation detection. *IEEE ICASSP*.
18. Cozzolino, D., et al. (2017). Recasting residual-based local descriptors. *IEEE TIFS*.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

19. Rossler, A., et al. (2019). FaceForensics++ dataset. *ICCV*.

20. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *CVPR*.

21. Wang, S., et al. (2020). CNN detection of GAN images. *CVPR*.

22. Gragnaniello, D., et al. (2021). GAN image detection. *Pattern Recognition Letters*.

23. Carion, N., et al. (2020). End-to-end object detection with transformers. *ECCV*.

24. Liu, Z., et al. (2021). Swin transformer. *ICCV*.

25. Coccomini, D., et al. (2022). Transformer networks for deepfake detection. *IEEE Access*.

26. Wodajo, D., &Atnafu, S. (2021). Deepfake detection using CNN-ViT hybrid. *Applied Sciences*.

27. Sabir, E., et al. (2019). Recurrent convolutional strategies. *CVPR Workshops*.

28. Dang, H., et al. (2020). Detection of AI-synthesized images. *IEEE Conference*.

29. Samek, W., et al. (2021). Explainable artificial intelligence. *IEEE Signal Processing Magazine*.

30. Ribeiro, M., et al. (2016). Why should I trust you? Explaining predictions. *KDD*.

31. Mahesh Ganji. (2025). Enhancing Oracle Cloud HR Reporting Through AI-Driven Automation. Journal of Science & Technology, 10(6), 28–36. https://doi.org/10.46243/jst.2025.v10.i06.pp28-36

32. Todupunuri, A. (2025). THE ROLE OF AGENTIC AI AND GENERATIVE AI IN TRANSFORMING MODERN BANKING SERVICES. American Journal of AI Cyber Computing Management, 5(3), 85–93. https://doi.org/10.64751/ajaccm.2025.v5.n3.pp85-93

33. Todupunuri, A. . (2024). Artificial Intelligence Ethics: Investigating Ethical Frameworks, Bias Mitigation, and Transparency in AI Systems to Ensure Responsible Deployment and Use of AI Technologies. International Journal of Innovative Research in Science,Engineering and Technology, 13(09), 1–14. https://doi.org/10.15680/ijirset.2024.1309002

34. Sushma Babburi. (2025). Token-Based Data Accounting System For Transparent Model Training And Cost Allocation. American Journal of AI Cyber Computing Management, 5(4), 463–474. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp463-474

35. Snigdha Gaddam. (2025). SOFTWARE STACK PREPARED FOR AI TRANSITIONING FROM MODULES TO MODELS. American Journal of AI Cyber Computing Management, 5(4), 451–462. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp451-462

36. Gaddam, S. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING.

37. Bajarang Bhagwat, V. (2023). Optimizing Payroll to General Ledger Reconciliation: Identifying Discrepancies and Enhancing Financial Accuracy. JOURNAL OF ADVANCE AND FUTURE RESEARCH, 1(4). https://doi.org/10.56975/jaafr.v1i4.501636

38. Srinivasa Kalyan Immadi. (2025). Harnessing Artificial Intelligence In Oracle Hcm: Revolutionising Workforce Management With Automation And Predictive Analytics. International Journal of Data Science and IoT Management System, 4(4), 7–13. https://doi.org/10.64751/ijdim.2025.v4.n4.pp7-13

39. S. M. K. P. (2025). Cryptography in iOS: A Study of Secure Data Storage and Communication Techniques. International Journal on Science and Technology, 16(1). https://doi.org/10.71097/ijsat.v16.i1.1403

40. Suhasnadh Reddy Veluru, Sai Teja Erukude, and Viswa Chaitanya Marella. 2025. Multimodal Detection of Fake Reviews using BERT and ResNet-50. In 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 877–882.

41. Cyril, H. P. (2025). Event-Driven Provisioning Architectures For Modern Telecom Networks: Overcoming Legacy Limitations And Enabling Autonomous 6g Operations. International Journal of Advanced Research in Computer Science, 16(6), 75–82. https://doi.org/10.26483/ijarcs.v16i6.7389

42. Jay Bharat Mehta. (2025). AUTONOMOUS PATCH VALIDATION FOR ZERO-DAY EXPLOITS IN ENTERPRISE CLOUDS. International Journal of Applied Mathematics, 38(4s), 1270–1285. https://doi.org/10.12732/ijam.v38i4s.685

43. Reddy, S. K. (2025). Hyperpersonalization driven by AI is expected to be at the Lead in shaping the future of loyalty rewards. Journal of Emerging Technologies and Innovative Research.

44. Reddy, S. K. R. (2021). Strengthening the Security of Loyalty Reward Systems: An In-Depth Analysis of Emerging Cyber Threats and Protection Mechanisms. Journal of Computational Analysis and Applications, 29(6).

45. Poojari, R. (2026). Privacy-Preserving Generative AI in Healthcare Systems Using Federated Learning Approaches. International Journal of Data Science and IoT Management System, 5(1), 78-88.

46. Uday Kumar Kalae. (2025). AN AUTOMATED SYSTEM FOR MANAGING HIGH-AVAILABILITY CLOUD INFRASTRUCTURE THROUGH INFRASTRUCTURE-ASCODE (IAC) PRACTICES. American Journal of AI Cyber Computing Management, 5(2), 42–50. https://doi.org/10.64751/ajaccm.2025.v5.n2.pp42-50

47. Saikumar, B. (2024). Optimizing Crew Scheduling and Absence Management using Microservices: Enhancing Reliability and Efficiency in Crew Management Systems. International Journal of Enhanced Research in Management & Computer Applications, 13(11), 50–55. https://doi.org/10.55948/ijermca.2024.0116

48. Saikumar, B. (2023). Enhancing Client Engagement through AI-Driven Real-Time Reporting and Automated Alerts. International Journal of Enhanced Research in Science, Technology & Engineering, 12(11), 111–117. https://doi.org/10.55948/ijerste.2023.1115