Volume 2 No 4, 2026
**Journal of Global Research in Multidisciplinary Studies (JGRMS)**
**Review Paper**
**Available online at** https://saanvipublications.com/journals/index.php/jgrms/index

# Hybrid CNN–Vision Transformer Architecture for Accurate Liver Cancer Diagnosis from Medical Imaging

Satyendra Sharma
*Department of Computer Science and Applications*
*ITM (SLS) Baroda University,*
Vadodara, Gujarat, India
Email: s.satya06@gmail.com

Pradeep Laxkar
*Department of Computer Science and Applications*
*ITM (SLS) Baroda University,*
Vadodara, Gujarat, India
pradeep.laxkar@gmail.com

***Abstract*—Detecting liver cancer early remains challenging because medical images can vary widely between patients, and differences in scan contrast are often subtle. This study describes a hybrid model that combines a CNN with a Vision Transformer, aiming to capture both fine, local image details and broader contextual information. In this setup, the CNN is used to focus on nearby visual signals such as edges and textures, while the transformer analyzes the full image to learn longer-range relationships between different regions. The method is evaluated on public datasets, including LiTS and TCGA-LIHC, with consistent preprocessing applied across all data. The reported accuracy is 94.8%, which is higher than the results from models using only a CNN or only a transformer. These findings indicate that leveraging both local and global features may lead to better performance in liver cancer detection**

***Keywords*—Hybrid Deep Learning, CNN, Vision Transformer, Liver Cancer, Medical Imaging, Feature Fusion, Transfer Learning**

.

## I. Introduction

Liver cancer is still one of the major causes of cancer-related deaths worldwide, and hepatocellular carcinoma (HCC) is the most common type [1][2][3]. Detecting it early can greatly improve survival, but finding tumors at an early stage is often difficult because medical images may show low contrast, tumors can have irregular shapes, and imaging features vary from one patient to another [4][5].

In clinical settings, diagnosis mainly depends on radiologists manually reviewing CT, MRI, and pathology images [6][7]. This process can be slow, and results may differ between readers. Deep learning methods, especially Convolutional Neural Networks (CNNs), have shown strong results in medical image analysis by learning layered visual features [8], but they are not always effective at capturing relationships between distant regions in an image.

Vision Transformers (ViTs) help with this by using self-attention to model broader context across the entire image [9]. Even so, ViTs often need very large datasets and may be less responsive to small, detailed local features. To address these issues, the study proposes a hybrid [10]model that combines CNN and transformer components, so it can learn both local spatial detail and global context, with the aim of improving diagnostic performance.

## II. Related Work

Deep learning has made automated medical image analysis more effective, particularly for detecting and classifying tumors [11][12]. In many studies, researchers use CNN-based models such as ResNet or DenseNet because they are good at learning local patterns within scans. A common drawback, however, is that these networks often concentrate on nearby regions, which may limit their ability to capture broader context across the entire image [13].

Vision Transformers aim to address this issue by using self-attention to link information from distant areas of an image [14]. In practice, though, they often depend on very large datasets and significant computing resources. For that reason, hybrid methods that combine CNN elements with transformer-like modules have become increasingly popular, since they can preserve fine-grained local detail while also incorporating wider contextual information.

## III. Proposed Methodology

The proposed approach follows a structured pipeline consisting of data collection, preprocessing, hybrid model development, feature fusion, and classification. Initially, medical imaging datasets are collected and prepared for training.

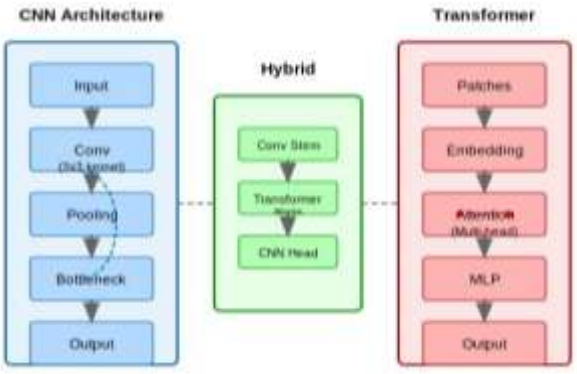


Fig. 1. Overall Workflow of the Proposed Liver Cancer Diagnosis Framework.

Pre-processing techniques are applied to maintain consistency and enhance data quality. These include normalization to standardize pixel intensity values, resizing images to a standard resolution of 224×224 pixels, noise reduction through filtering techniques, and data augmentation to increase dataset diversity. The overall workflow of the proposed approach is illustrated in Fig. 1.

The hybrid model combines CNN and transformer components to extract complementary features. The CNN module processes the input image to capture local spatial features, while the transformer module processes image patches to learn global dependencies. The outputs of both components are combined through a feature fusion mechanism, which combines local and global representations into a unified feature space. The structure of the CNN module employed for feature extraction is shown in Fig. 2[15]

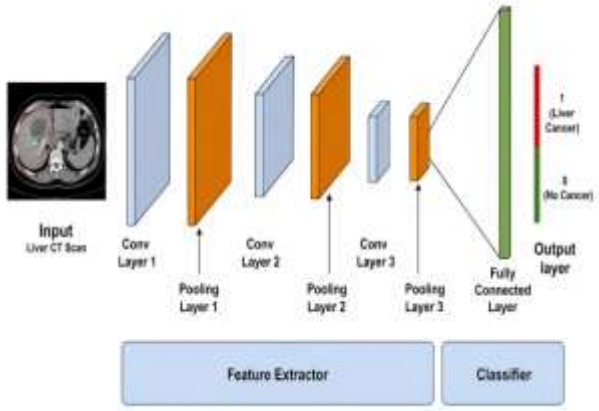


Fig. 2. CNN Architecture Illustrating Hierarchical Feature Extraction.

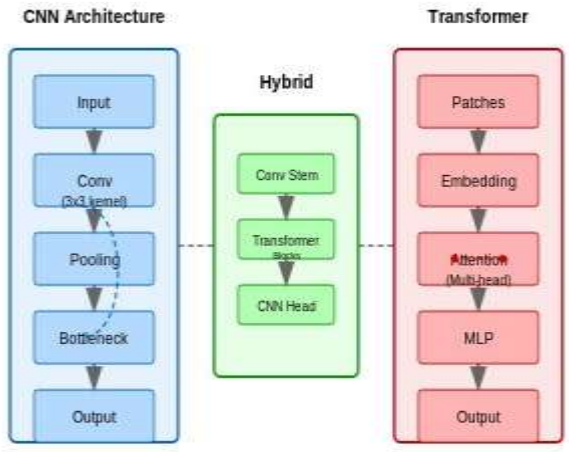


Fig. 3. Overall Workflow of the Proposed Liver Cancer Diagnosis Framework.

The hybrid model combines CNN and transformer components to extract complementary features. The CNN module processes the input image to capture local spatial features, while the transformer module processes image patches to learn global dependencies. The outputs of both components are combined through a feature fusion mechanism, which combines local and global representations into a unified feature space. The structure of the CNN module employed for feature extraction is shown in Fig. 2

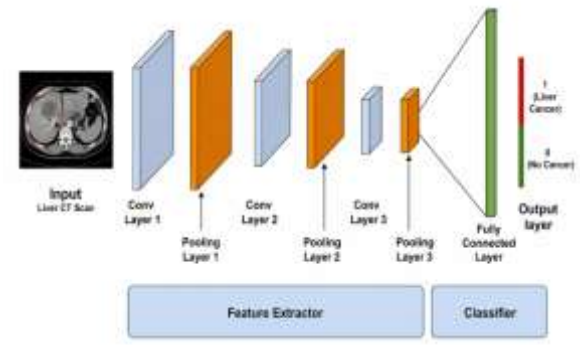


Fig. 4. CNN architecture illustrating hierarchical feature extraction.

The transformer-based feature extraction process is illustrated in Fig. 3. The Vision Transformer partitions the input image into fixed-size patches and processes them using self-attention mechanisms to capture global contextual relationships[16].

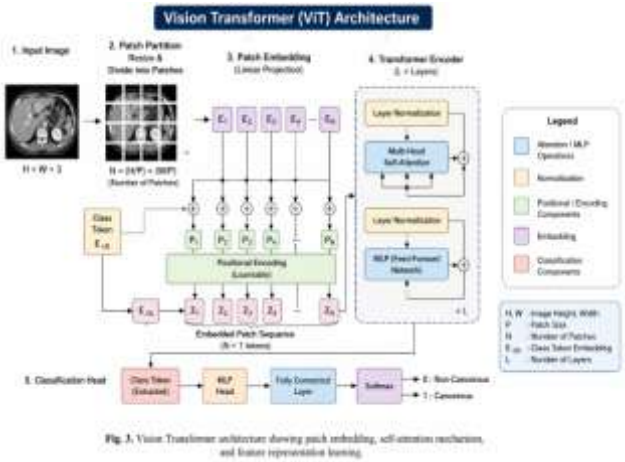


Fig. 5. Vision Transformer Architecture Showing Patch Embedding and Self-Attention Mechanism.

Mathematically, the hybrid feature representation can be expressed as:$F_{hybrid} = \alpha F_{CNN} + \beta F_{ViT}$ (3.1)

Where $F_{CNN}$ represents local features, $F_{ViT}$ represents global features, and $\alpha$, $\beta$ are weighting factors.

The complete hybrid CNN–Vision[17] Transformer architecture with feature fusion is illustrated in Fig. 4.

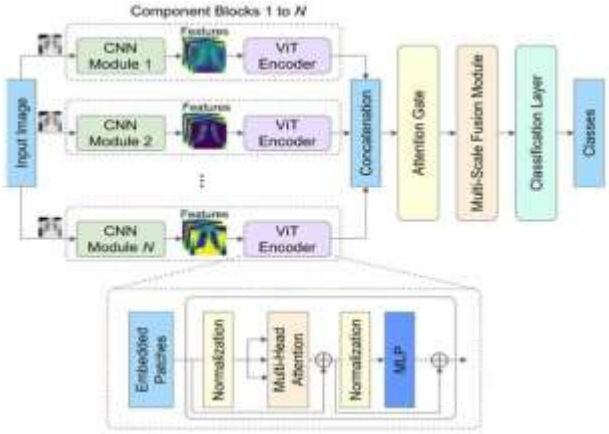


Fig. 6. Architecture of the proposed hybrid CNN–Vision Transformer model.

As illustrated in Fig. 4, input medical images are processed using multiple CNN modules to extract local spatial features. These features are subsequently provided to Vision Transformer encoders, where global contextual relationships are captured using self-attention mechanisms. Extracted features from different branches are concatenated and refined via an attention gate and multi-scale fusion module. Finally, the fused representation is processed through a classification layer predict liver cancer presence or absence.

By combining CNNs with transformer parts, the model can pick up detailed spatial patterns while also capturing longer-range relationships, which can help it, make more accurate diagnoses.

## IV. Dataset and Preprocessing

The approach is evaluated on public datasets to make the results easier to verify and more likely to generalize. In this study, the LiTS dataset[18] is used to detect tumors in CT scans, while TCGA-LIHC is used to classify histopathology images. Before training, the images undergo pre-processing: pixel values are normalized, images are resized to 224×224, and noise is reduced [19][20]. To introduce more variation and support better generalization, data augmentation is also applied, such as rotation, flipping, and scaling. Together, these steps aim to standardize the inputs and help limit over fitting.

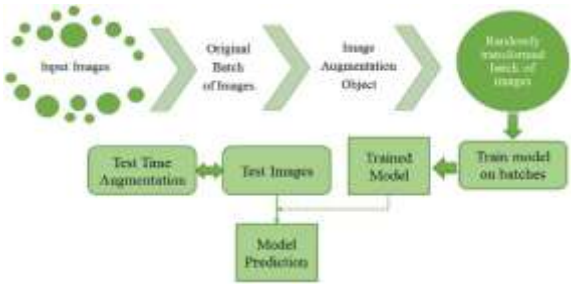


Fig. 7. Workflow of image augmentation and model prediction.

Key characteristics of the datasets are summarized in Table I.

TABLE I. Dataset characteristics for model training and evaluation

| Dataset | Modality | Cases | Purpose |
|---|---|---|---|
| LiTS | CT | 131 | Tumor Detection |
| TCGA-LIHC | Histopathology | ~350 | Classification |

Table II has a list of everything we did to get the data ready.

TABLE II. Preprocessing steps applied to medical imaging data

| Step | Description |
|---|---|
| Normalization | Pixel intensity scaling |
| Resizing | 224 × 224 resolution |
| Augmentation | Rotation, flipping, scaling |
| Noise Reduction | Gaussian filtering |

Following dataset preparation and preprocessing, the model is optimized under the experimental setup described below.

## V. Experimental Setup

Training uses the Adam optimizer with an initial learning rate of 0.0001. The model is trained for 50 epochs with a batch size of 32. All experiments run on GPU-enabled systems to improve speed. The setup is intended to support reproducibility and allow fair comparison with baseline models.

## VI. Results and Discussion

The proposed hybrid CNN–Vision Transformer model is compared against a CNN-only model and a transformer-only model. The proposed hybrid approach appears to deliver better overall performance, achieving 94.8% accuracy and showing stronger sensitivity and specificity.[21]

The improvement appears to come from the model's capacity to learn detailed spatial cues as well as broader context. In practical terms, the CNN component focuses on local signals like edges and textures, while the transformer component is used to capture relationships between regions that may be far apart in the image. When these two parts are combined, the model can represent complex tumor structures more effectively.

Higher sensitivity matters in medical diagnosis because it lowers the chance of overlooking real cancer cases. The results also seem to generalize consistently across different datasets, which suggests the approach is reasonably robust. A comparative overview is provided in Table III.

TABLE III. Comparative performance of CNN, ViT, and hybrid models

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| CNN | 92.10% | 89.70% | 94.50% |
| ViT | 91.50% | 88.20% | 93.80% |
| Hybrid | 94.80% | 92.50% | 96.30% |

The results indicate that the proposed hybrid model provides consistent improvements across multiple evaluation metrics, confirming the effectiveness of integrating CNN and transformer-based feature extraction[22].

The model performance is further illustrated using ROC curves and confusion matrix, as illustrated in Fig. 6 and Fig. 7

To assess model performance, Receiver Operating Characteristic (ROC) curves are evaluated, illustrating the relationship between true positive rate and false positive rate across different thresholds[23]

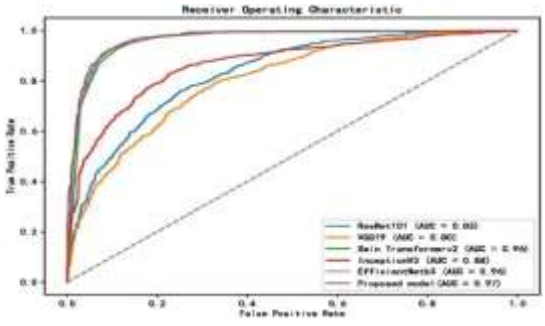


Fig. 8. Receiver Operating Characteristic (ROC) curves comparing the proposed hybrid CNN–Vision Transformer model with baseline architectures.

As illustrated in Fig. 6, the proposed hybrid approach achieves the highest Area Under the Curve (AUC = 0.97), outperforming baseline models including ResNet101 (AUC = 0.83), VGG19 (AUC = 0.80), InceptionV3 (AUC = 0.88), and EfficientNetB3 (AUC = 0.96). A higher AUC value reflects improved classification capability and improved discrimination between cancerous and non-cancerous cases. This performance improvement highlights the effectiveness of

combining CNN-based local feature extraction with transformer-based global contextual modeling[24].

These results indicate that hybrid architectures significantly improve diagnostic accuracy by leveraging complementary feature representations[25].

To further analyze classification performance, the confusion matrix is presented in Fig. 7.[26]

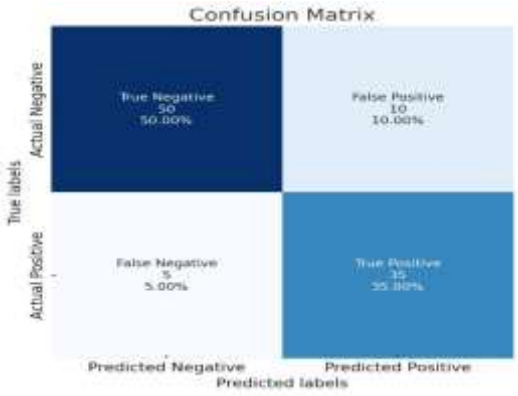


Fig. 9. Confusion matrix of the proposed hybrid model.

The confusion matrix indicates that the proposed model achieves high true positive and true negative rates, thereby reducing misclassification errors[27].

## VII. Ablation Study

We ran an ablation study to see what each part of the proposed model actually adds. The results suggest the CNN is good at picking up local spatial details, while the transformer helps capture broader context across the input. When we combine the two, the model performs better overall, which supports the idea behind using this hybrid setup. The numbers showing how each component affects performance are listed in Table 4[28].

## VIII. Ablation Results

The impact of different architectural components is evaluated in Table 4.

TABLE IV. Ablation study evaluating the contribution of individual components in the proposed hybrid CNN–Vision Transformer model.

| Configuration | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| CNN Only | 92.10% | 89.70% | 94.50% |
| ViT Only | 91.50% | 88.20% | 93.80% |
| CNN + ViT (No Fusion Optimization) | 93.20% | 90.80% | 95.10% |
| Proposed Hybrid Model | 94.80% | 92.50% | 96.30% |

## IX. Conclusion

This study describes a hybrid CNN and Vision Transformer model for diagnosing liver cancer using medical imaging. The CNN is used to pick up local patterns in the images, while the transformer helps capture broader context across the full scan, which together supports stronger classification results. In experiments, the model shows higher accuracy, sensitivity, and specificity than the baseline methods used for comparison. The next stage of work will concentrate on improving computational efficiency and adapting the approach to multimodal medical data so it is more suitable for real-world clinical use.

## References

[1] M. R. Ahmed, “Deep Learning in Medical Imaging: A Comprehensive Review of Techniques, Challenges, and Future Directions,” vol. 12, no. 12, 2025, doi: 10.4236/oalib.1114497.

[2] B. Sahu *et al.*, “Harnessing TLBO-Enhanced Cheetah Optimizer for Optimal Feature Selection in Cancer Data,” *Comput. Model. Eng. Sci.*, pp. 1–26, 2025, doi: 10.32604/cmes.2025.069618.

[3] R. A. Jahromi, A. Abootalebzadeh, M. Abbasi, and A. Sharafkhah, “Artificial Intelligence for Liver Cancer Diagnosis: Integrating Image Analysis, Algorithms, and Clinical Applications,” vol. 2, no. 10, pp. 42–62, 2025, doi: 10.61882/ist.202502.10.04.

[4] R. Kant, R. Rao Thallada, B. Pandey, and P. Srivastava, “AI-Based Cybersecurity in Healthcare: A Data-Driven, Governance-Aware Framework for Secure Clinical Systems,” in *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*, 2026, pp. 1–5. doi: 10.1109/ICAIC67076.2026.11395836.

[5] S. Dharmavaram and P. Bhanushali, “Machine Intelligence-Driven Forecasting for ED Triage and Dynamic Hospital Patient Routing,” *medRxiv*, 2026, doi: 10.64898/2026.02.18.26346566.

[6] C. Tayal and S. Murumkar, “Patient Identity Protection and Duplicate Record Prevention in Electronic Health Record (EHR) Systems,” in *2026 18th International Conference on Knowledge and Smart Technology (KST)*, 2026, pp. 458–464. doi: 10.1109/KST67832.2026.11431915.

[7] R. Snehamrutha, “Development and In-Vitro Evaluation of a Sustained-Release Transdermal Patch for Improved Patient Adherence,” *ESP Int. J. Adv. Sci. Technol.*, vol. 2, no. 4, pp. 34–43, 2024, doi: 10.56472/25839233/IJAST-V2I4P105.

[8] S. K. Anumula, “Brain-Inspired Computing: Revolutionizing Medical Devices Through Neuromorphic Engineering, A Review Of Different Options,” *Int. J. Comput. Sci. Eng. Res. Dev.*, vol. 15, no. 2, pp. 88–105, 2025, doi: 10.63519/IJCSERD_15_02_007.

[9] S. Mahmud, “AI and Data Analytics for Enhancing Home Healthcare: Optimizing Patient Outcomes and Resource Allocation,” *Front. Appl. Eng. Technol.*, vol. 2, no. 1, pp. 23–100, 2025, doi: 10.70937/faet.v2i01.61.

[10] E. Othman, M. Mahmoud, and H. Dhahri, “Automatic Detection of Liver Cancer Using Hybrid Pre-Trained Models,” 2022, doi: 10.3390/s22145429.

[11] T. P. Patel, A. K. Elengovan, V. Ranganathan, M. Parikh, and D. Kole, “Self-Healing AI Systems Using Multi-Agent Learning,” in *2026 International Seminar on Intelligent Business and Edge-Computing Research (ISIBER)*, 2026, pp. 7–12. doi: 10.1109/ISIBER68248.2026.11470173.

[12] A. V. S. R. Dantuluri and S. Kumar, “A Governance-Driven, Real-World Data-Calibrated Health Informatics Framework for Longitudinal Utilization Forecasting in Oncology and Complex Chronic Conditions,” *medRxiv*, 2026, doi: 10.64898/2026.02.23.26346919.

[13] J. A. Kachhia, “Healthcare Predictive Analytics based on Machine Learning Techniques for Identifying Cardiovascular Risks Screening,” *Int. J. Curr. Eng. Technol.*, vol. 13, no. 6, pp. 342–635, 2026, doi: 10.14741/ijcet/v.13.6.17.

[14] A. Warrier, “Real-Time AI Integration Architectures for HIPAA-Compliant Healthcare Data Interoperability,” in *International Journal of Emerging Trends in Computer Science and Information Technology*, Eureka Vision Publication, 2025, pp. 74–81. doi: 10.63282/3050-9246/WCAI25-128.

[15] S. N. Rao and C. D. V Subba, “Liver Tumor Segmentation and Classification Using Deep Learning Methods,” 2024, doi: 10.1109/ICDCOT61034.2024.10515854.

[16] F. Shamshad, S. Khan, S. W. Zamir, and M. H. Khan, “Transformers in medical imaging: A survey,” 2023, doi: 10.1016/j.media.2023.102802.

[17] X. Liu, Y. Hu, and J. Chen, “Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron,” 2023.

[18] P. Bilic, P. Christ, and H. B. L, “The Liver Tumor Segmentation Benchmark (LiTS),” vol. 84, 2023, doi:

 

10.1016/j.media.2022.102680.

[19] A. K. Padhy, N. Seshagiri, V. Soni, A. K. Elengovan, G. Babu Thokala, and M. Kumari, "AI-Enabled Autonomous Data Preprocessing: A Scalable Architecture for Intelligent Machine Learning Pipeline Management," in *2026 6th International Conference on Image Processing and Capsule Networks (ICIPCN)*, 2026, pp. 1318–1323. doi: 10.1109/ICIPCN67432.2026.11438503.

[20] R. V. S. S. B. R, A. M, S. Sharhrah, K. Keerthana, and S. Senthil, "A Personalized Federated Multi-Task Learning Framework for Multi-Modal COVID-19 Diagnosis at the Edge," in *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, 2025, pp. 1–6. doi: 10.1109/NMITCON65824.2025.11188318.

[21] Y. Li and P. Song, "Review of transfer learning in medical image classification," *J. Image Graph.*, vol. 27, no. 3, pp. 672–686, 2022, doi: 10.11834/jig.210814.

[22] K. Borys, Y. A. Schmitt', and M. Nauta, "Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches," 2023, doi: 10.1016/j.ejrad.2023.110787.

[23] C. AMADI and A. OJO, "Building Trustworthy AI in Healthcare," 2025.

[24] S. Kumar, "Evaluation Metrics For Classification Model," *2025*.

[25] S. A. H. Shah *et al.*, "Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning," *J. Imaging*, vol. 10, no. 12, p. 332, Dec. 2024, doi: 10.3390/jimaging10120332.

[26] Z. Vujovic, "Classification Model Evaluation Metrics," 2021, doi: 10.14569/IJACSA.2021.0120670.

[27] Z. Liu *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," 2021.

[28] M. Gunay, Ö. Yıldırım, and Y. Demir, "Deep Learning: Evolution, Innovations, And Applications In The Last Decade," 2025, doi: 10.51477/mejs.1640908.

 