The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
20(3): S.I (3), 1311-1322, 2025 www.thebioscan.com

EXPLAINABLE TWO-STAGE VISION TRANSFORMER FRAMEWORK FOR CORAL REEF DISEASE DETECTION AND INTERPRETATION

M. H. Ibrahim[1], Dr.S. Muruganantham[2]

[1]Research Scholar (Reg No: 18121192161001), Department of Computer Science, S.T. Hindu College, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamilnadu, India.
[2]Associate professor and Research Supervisor, Department of Computer Science, S.T. Hindu College, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamilnadu, India.

DOI: 10.63001/tbs.2025.v20.i03.S.I(3).pp1311-1322

KEYWORDS:
Coral reef classification, Deep learning, Explainable artificial intelligence, Grad-CAM, Vision transformer

Received on:

05-09-2025

Accepted on:

01-10-2025

Published on:

04-11-2025

ABSTRACT

Coral reefs are increasingly threatened by climate change and diseases like white band disease. Accurate monitoring of coral health is essential for effective conservation. Several machine learning and deep learning models have been developed for coral reef monitoring, achieving high accuracy. However, most existing models lack interpretability and do not provide insight into the reasoning behind their predictions. To address this limitation, this paper proposes a novel architecture for coral reef type classification and white band disease detection using a two-stage Vision Transformer (ViT) framework combined with Explainable Artificial Intelligence (XAI) technique. Coral reef images undergo preprocessing to enhance quality, followed by augmentation to expand the dataset and improve model robustness. These processed images are fed into the two-stage ViT framework for feature extraction and classification. In the first stage, the model identifies the type of coral reef. In the second stage, the original image is analyzed together with the stage one output to detect the presence of white band disease. Performance of the proposed model is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. Grad-CAM visualization is employed to highlight the regions influencing the model's decisions, providing interpretability and increasing trust in predictions. Experimental results demonstrate that the proposed framework not only accurately classifies coral reef types but also effectively detects white band disease with higher performance compared to existing methods. The integration of XAI and two-stage ViT architecture enables both precise predictions and interpretable results, making the framework a valuable tool for coral reef monitoring and conservation efforts.

1. INTRODUCTION

Coral reefs are among the most diverse and productive ecosystems on earth. They are often called as rainforest of the sea. They provide essential habitats for approximately 25% of marine species, including over 4000 fish species and contributes to coastal protection, nutrient cycling, fisheries, and tourism [1][2]. The vibrant structure of coral reefs is maintained through the symbiotic relationship between corals and zooxanthellae algae, which facilitate calcium carbonate skeleton formation. However, rising sea surface temperatures due to global warming cause coral

bleaching, leading to the expulsion of these algae and a loss of coral vitality. Bleaching events have significant ecological and economic consequences, including biodiversity loss, reduced fisheries productivity, and tourism decline [3].

Ocean acidification, driven by elevated CO_2 levels, further hinders the calcification process, compromising coral growth and resilience. Coral reefs also host organisms that produce bioactive compounds with potential pharmaceutical applications. Continuous monitoring of coral reefs is critical to mitigate the impacts of climate change and human activities, including overfishing, pollution, and coastal development [4][5]. However, underwater imagery analysis is challenging due to noise, light attenuation, and other environmental artifacts. Recent advances in underwater imaging, remote sensing, and artificial intelligence (AI) have enabled more effective coral reef monitoring. Machine Learning models combined with geospatial analysis tools, allow for automated detection, classification, and prediction of coral health and environmental impacts. By integrating AI with geospatial and remote sensing data, researchers can better understand reef dynamics, detect diseases, and implement targeted conservation strategies. This holistic approach enhances large-scale, long-term coral reef monitoring, supporting both ecological preservation and sustainable management of these vital marine ecosystems.

Deep Learning (DL) models have achieved remarkable success in detecting coral reef diseases from images and videos [6][7][8][14]. However, most existing DL based models act as black boxes. They provide accurate prediction without explaining the underlying reasoning. This lack of interpretability limits trust and adaption by marine scientists in real-world coral reef monitoring applications. To address such an issue, this paper proposes an Explainable Artificial Intelligence supported two-stage Vision Transformer (ViT) framework that combines self-attention mechanisms and explainability modules for reliable and interpretable coral reef type classification and white band disease detection. Unlike earlier coral disease detection systems that focus only on accuracy, the proposed framework introduces transparency and explainability into the decision- making process. The novelty of the proposed method is that the integration of XAI into a two-stage ViT. Design of visual interpretability interface that bridges the gap between computational prediction and ecological understanding. The key contributions of this are as follows:

- To develop DL model for accurate detection and classification of coral reef types and disease
- To integrate attention based explainability mechanisms that visualize disease affected coral regions
- To improve interpretability via Grad-CAM visualizations
- To evaluate the accuracy and interpretability of the model against DL based models.
- To the best of authors' knowledge, this is the first attempt to develop a transparent, explainable, and ecologically validated DL model for coral disease detection.

The rest of the paper is organized as follows: Section 2 provides a brief review of related works. Section 3 explains the proposed methodology. Simulation results are presented in section 4. Section 5 concludes the paper.

2. LITERATURE ANALYSIS

Jamil et al. [4] proposed a bag-of-features based method to detect and localize bleached corals. Authors extracted hand-crafted features like Local Binary pattern

(LBP), Histogram of Oriented Gradients (HoG), and Gray Level Co-occurrence Matrix (GLCM) from the underwater images. They introduced a bleached coral positioning algorithm to precisely locate bleached corals. Support Vector Machine (SVM) with various kernels was used to detect bleached corals.

Fawad et al. [5] focused on developing a Robust Localization using Bag-of-Hybrid Visual Features (RL-BoHVF) method to classify the coral images. The model utilized both deep features extracted by AlexNet and hand-crafted features to improve classification accuracy. Results demonstrated that the combined features enhanced classification accuracy. However, the model was evaluated on a limited dataset, which restricts its generalizability to other reef regions.

Ibrahim and Sathik [6] designed a hybrid ZeNet and VGG19 network to detect white band corals from videos. The framework extracted features using AlexNet and hybrid net to improve invariance and classification accuracy. Results showed that the hybrid method detected the diseased corals with classification accuracy of 98.02%.

Chowdhury et al. [7] presented a comprehensive review of coral disease detection methods. Author discussed about the datasets and reviewed ML and DL machine learning methods used to detect diseased corals. They also mentioned the merits and demerits of earlier methods.

Aldhahri et al. [8] implemented YOLOv8 and YOLOv9 object detection models to classify coral images into three categories: healthy, bleached, and dead corals. The models were trained and evaluated on augmented dataset to improved detection accuracy. Results showed that the YOLOv9 model attained higher precision compared to YOLOv8.

Wang et al. [9] presented a Multi-Local Perception Network (ML-Net) for the classification of healthy and bleached corals. The model adopted a multi-branch local adaptive block to extract detailed image features. Residual structures in the shallow layers transmitted local texture and color information to deeper layers, which helped retain local details ML-Net yielded an accuracy of 86.35% and outperformed ResNet and ConvNext.

Karthik et al. [10] examined the potential of ML methods for automatic classification of coral reefs from underwater images. Authors stresses the need for monitoring and conservation. ML methods were applied to solve issues like coral bleaching. They provided insight on reef health and guided effective reef management.

Trudeau et al. [11] presented a comparative study of ML models, including Convolutional Neural Network (CNN) and Binary Logistic Regression (BLR) for determining coral reef types. The CNN outperformed the BLR in t accuracy and F1-score.

Ogidi and Sah [12] used pre-trained DL models including ResNet-152, EfficientNet-B0, and DenseNet-121 to detect coral diseases. The models were trained and evaluated on BC-HC and CRHI datasets. Results demonstrated that the EfficientNEt-B0 attained better results than other networks.

Firdous and Sabena [13] introduced a method to improve coral species classification from underwater images. Authors integrated Tri Convolutional Deep Autoencoders (TRI-CADE) and Sparse Deep Autoencoders (SPDAE) to classify the images. Results showed that the achieved model yielded superior outcomes to other models.

3. MATERIALS AND METHODS

DL based models for image classification have rapidly moved beyond CNNs to embrace transformer-based architecture. ViTs have demonstrated strong representation capabilities in a variety of image classification tasks due to their ability to model long-range dependencies

and global context. This study proposes an XAI-ViT framework for coral type classification and white band disease detection. The model integrates transformer-based feature learning with explainability to ensure both accuracy and interpretability. Overall process of the proposed framework is shown in Figure 1. The model comprises four major phases: data preparation, two-stage ViT, explainability, and performance valuation.

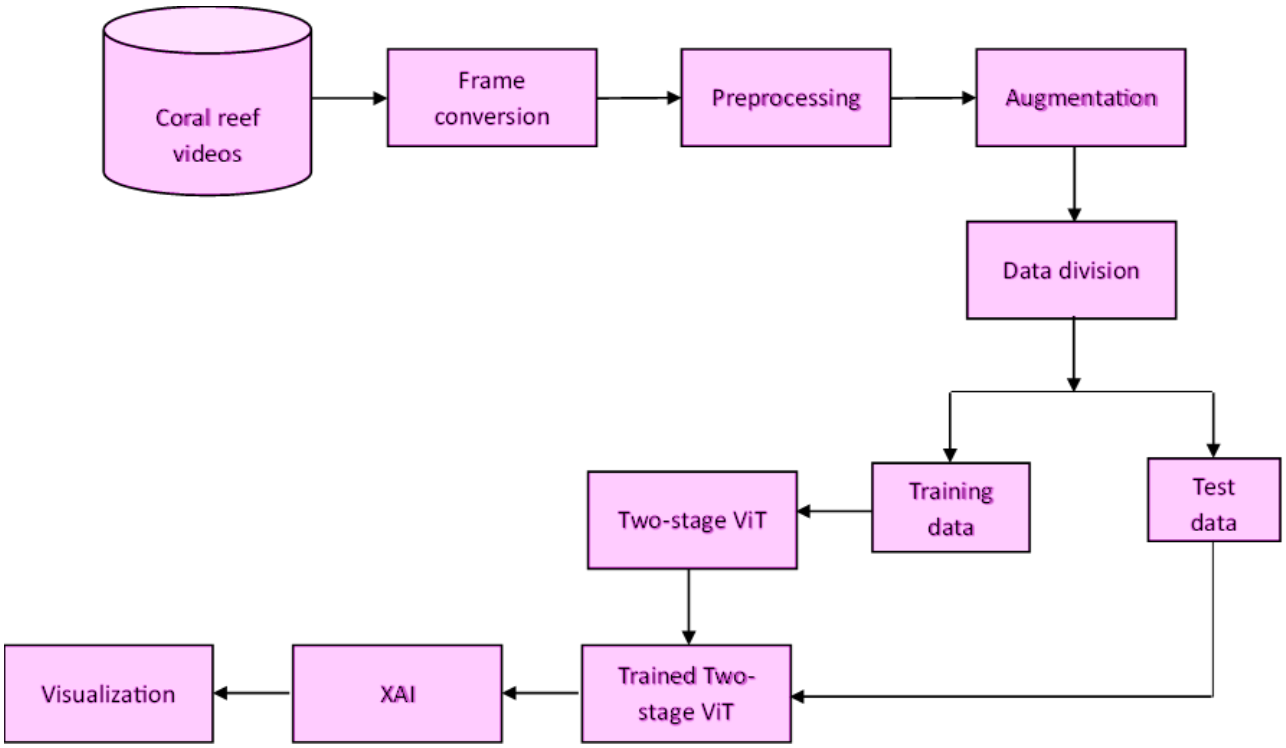


Figure 1 Pipeline of the proposed methodology for coral disease detection

3.1 Dataset preparation

This study employs two distinct datasets to train and evaluate a two-stage ViT for coral health monitoring. The first data is gathered from Mendeley, consists of 1512 images representing a wide variety of coral reef species. This dataset is used to train the two-stage ViT to classify different coral types. To ensure effective training and evaluation, 80% of the images are used for training, while the remaining 20% are reserved for testing. The second data consist of a 3.12hour underwater video sources from the official BBC Earth You Tube channel. From this video, 22,462 frames are extracted at a rate of two frames per second. 3000 frames are selected for training (80%) and testing (20%). By combining these datasets, the proposed XAI-ViT framework benefits from both structured image and video data. The first data allows network to learn distinguishing features of different coral species, while the second data trains the model to recognize disease patterns in real-world underwater conditions.

Underwater images often suffer from low contrast, noise, and color degradation. To mitigate these effects, the following preprocessing methods are applied: Color correction is performed using Contrast Limited Adaptive Histogram

Equalization (CLAHE) to improve color balance followed by bilateral filtering to reduce noise. Pixels values are normalized and all images are resized to match ViT input requirements. Data augmentation including random rotations, flips, zoom, and brightness adjustments is applied to improve model generalization. Figure 2 shows sample augmented images.

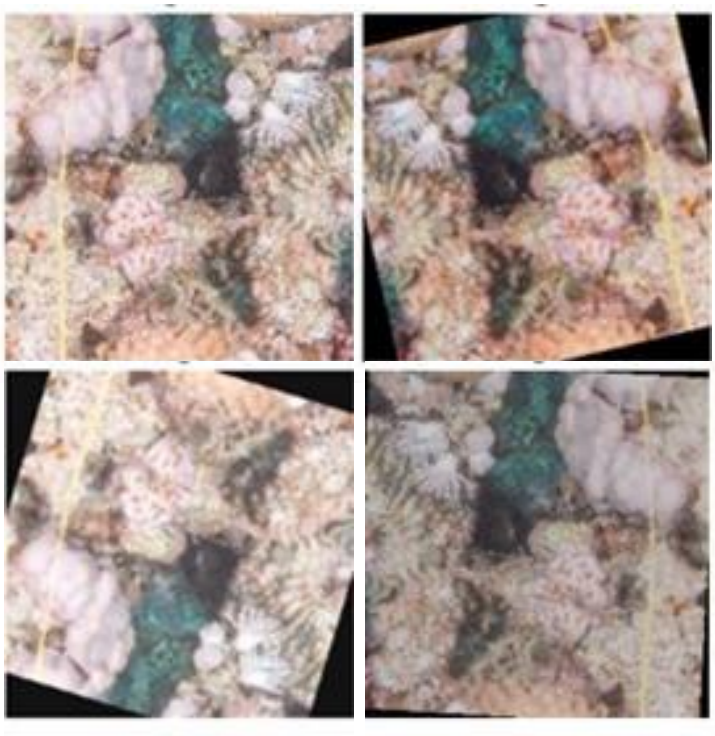
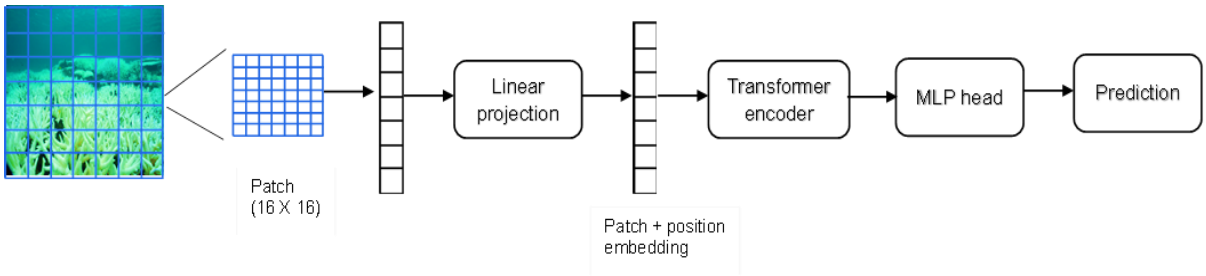


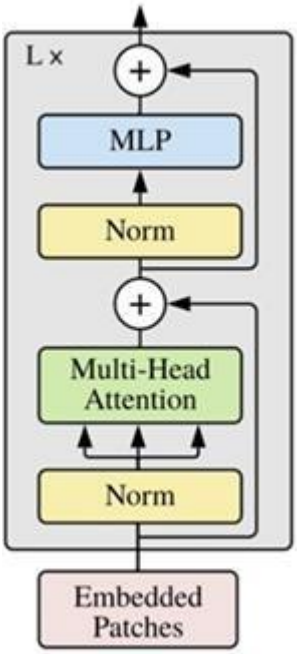
Figure 2 Pre-processed and augmented images

3.2 Two-stage ViT architecture

Unlike traditional CNNs that classify coral diseases directly, the proposed framework divides the task into two stages to improve accuracy and interpretability. The first stage aims to recognize the type of coral present in the image. The ViT is an encoder-decoder architecture based on multi head self-attention. Encoder transforms input data into lower-dimensional vector representation, while decoder process feature vector to generate output. Figure 3 shows structure of ViT. Input image is split into 16 x 16 patches and then reshaped into a vector of dimension (1,256). Each patch is projected into a new space of dimension. This projector vector is the feature vector, as illustrated in Figure 3. These feature vectors are passed through encoder and then final prediction is made by projecting final embeddings via multilayer perceptron.



(a)



(b)

Figure 3 ViT (a) Main architecture of ViT (b) Transformer encoder

Encoder block comprises of layer normalization, multi-head self-attention and MLP. Layer normalization enhances learning speed. The attention mechanism transforms the representation of input sequence into contextualized one. Equation (1) represents the formula for calculating enriched embedding. Attention scores can be computed by using Equation (2), The softmax function converts attention score into a probability distribution.

$$Enriched\ embedding(Q,K,V) = softmax\ (\frac{QK^T}{\sqrt{d_k}})\ V \quad (1)$$

$$Attention\ score(Q,K) = softmax\ (\frac{QK^T}{\sqrt{d_k}}) \quad (2)$$

In coral type classification, multi-head attention mechanism in ViT captures long-range dependencies across textures and color patterns. These extracted attributes are used to determine type of coral. The second stage of the framework identifies whether the coral exhibits white band disease or not. To achieve this, second ViT uses both coral type prediction from stage 1 and original coral image as inputs. This transformer encoder in ViT-2 focuses on disease specific cues to identify the presence of white band disease. The

softmax head output probabilities across two values. This two-stage approach enhances interpretability and reduces inter-class confusion by allowing each model to specialize in a different visual subtask.

3.3 Explainability

To address the opacity of transformer models, XAI module is incorporated. Gradient-Weighted Class Activation Mapping (Grad-CAM) utilized to visualize which regions of coral image contribute most to model's prediction. It calculates the gradient of target class score with respect to feature maps of the transformer encoder layer, thereby highlighting discriminative regions.

4. EXPERIMENTAL RESULTS

Explainable two-stage ViT was implemented and validated using Pytorch, which is an open-source DL platform that offers flexibility and ease of use for research. Performance and interpretability of the proposed framework is assessed by computing the following metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$F1-\text{score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision}+\text{Recall})} \quad (7)$$

Where, TP-number of images that the proposed model predicts as positive, TN-number of images that the proposed model predicts as negative, FP- number of images that the proposed model incorrectly predicts as positive, FN-number of images that the proposed model incorrectly predicts as negative.

4.1 Results analysis on Coral reef type classification

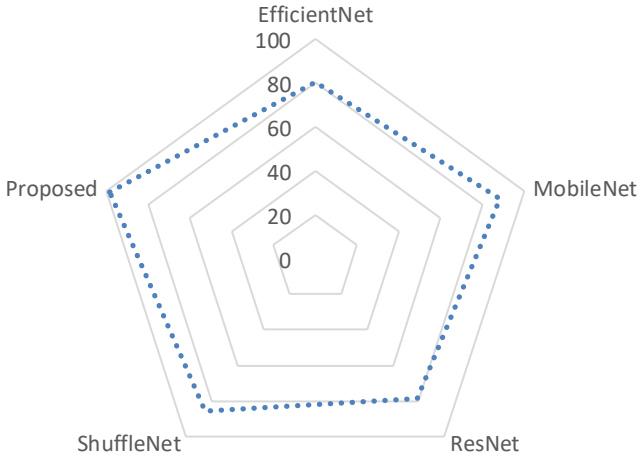
Table 1 Performance comparison of the proposed model with other deep learning models for coral reef type classification

Models	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-score
EfficientNet	80.33	83.67	77	78.44	80.97
MobileNet	88.33	91	85.67	86.39	88.64
ResNet	78.67	81.67	75.67	77.04	79.29
ShuffleNet	85.67	87.02	84.33	84.74	85.86
Proposed	98.37	94	90.42	91.58	92.76

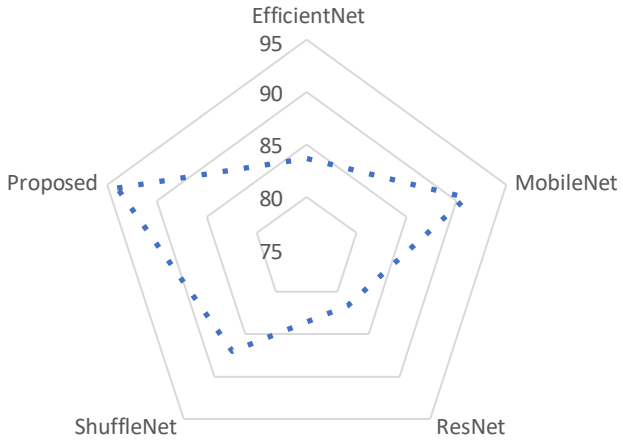
Table 1 compares the performance of the proposed framework against four DL methods including EfficientNet, MobileNet, ShuffleNet, and ResNet. The EfficientNet attained an accuracy of 80.33% and recall of 83.67%. Its moderate performance suggests that while the compound scaling strategy of EfficientNet balances depth and width, the network may have limited capacity to extract complex coral textures present in the dataset. The ResNet achieved an accuracy, recall, specificity, and F1-score of 78.67%, 81.67%, 75.67%, and 79.29%, respectively. These results show that ResNet's residual connections enhance gradient flow but may

not capture subtle spatial dependencies. ShuffleNet yielded an accuracy of 85.67% and F1-score of 85.86%. Both models showed lower discriminative capability compared to advanced DL models. MobileNet demonstrated better performance by achieving 88.33% accuracy, 91% recall, and 88.64% F1-score. It's depth wise separable convolutions enable efficient feature extraction while maintaining a balance between precision and recall. The two-stage ViT outperformed all other DL models by achieving an impressive accuracy of 98.37%, 94% recall, 90.42% specificity, and 92.76% F1-score. This performance demonstrated the superiority of ViT model for coral classification tasks. The ViT's self-attention mechanism allows it to capture global relationships across the entire image. Two-stage design enhances hierarchical feature learning, enabling the model to generalize across multiple coral types with minimal misclassification.

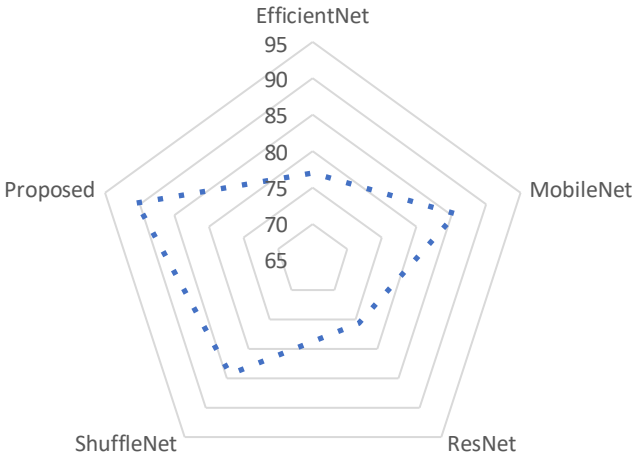
4.2 Result analysis on white band disease detection



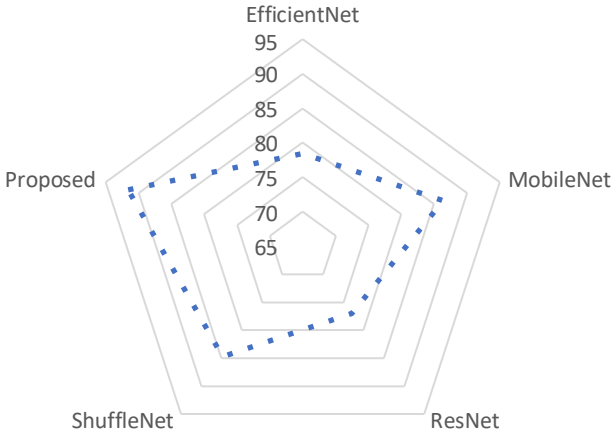
(a) Accuracy



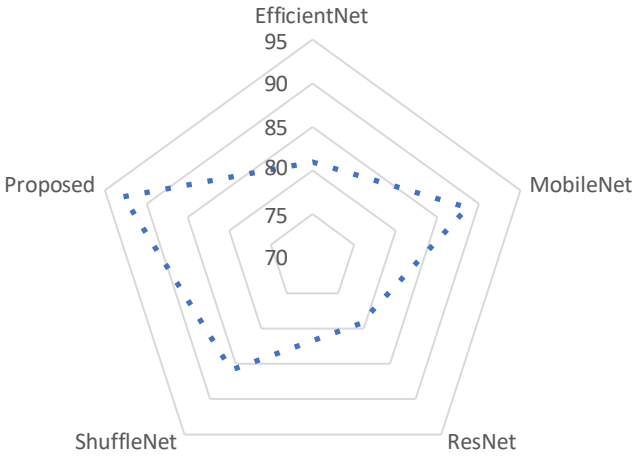
(b) Recall



(c) Specificity



(d) Precision



(e) F1-score

Figure 4 Comparison of the proposed model's performance with other deep learning models for White Band Disease detection

Figure 4 Illustrates The comparative performance of several deep learning models for detecting white band disease in coral images shows a progressive improvement from conventional CNN architectures to transformer-based approaches. The EfficientNet model achieved an accuracy of 80.33%, recall of 83.67%, and F1-score of 80.97%, capturing general disease features but with limited specificity. MobileNet improved performance to 88.33% accuracy and 88.64% F1-score, owing to efficient depth wise separable convolutions that enhance fine-grained feature extraction. ResNet recorded lower accuracy (78.67%) and F1-score (79.29%), indicating challenges in handling complex coral patterns, while ShuffleNet achieved 85.67% accuracy and 85.86% F1-score through its efficient channel shuffling mechanism. In contrast, the proposed Two-Stage Vision Transformer (ViT) demonstrated superior performance with an accuracy of 98.37%, recall of 94%, and F1-score of 92.76%. Its hybrid design integrates convolutional layers with multi-head self-attention, enabling the model to capture both local spatial patterns and global contextual relationships among coral fragments. The significant improvement in accuracy and F1-score confirms that attention-based architectures offer deeper feature understanding and better generalization than traditional CNN models. Therefore, the Two-Stage ViT conclusively stands out as the most effective and robust model for accurate detection of white band disease in coral images.

4.3 Explainability analysis

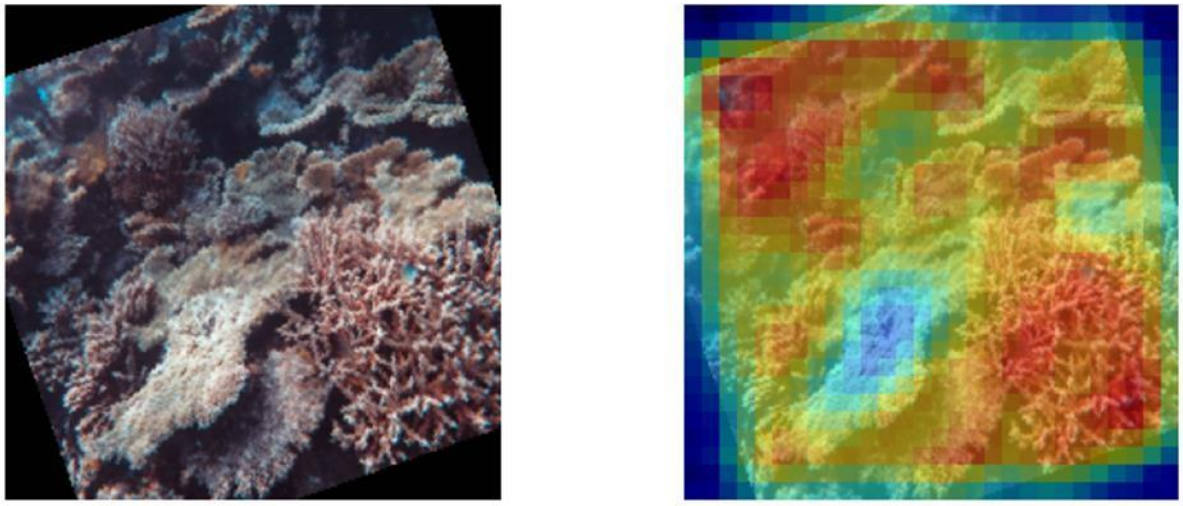


Figure 5 Grad-CAM visualization

Figure 5 shows the Grad-CAM visualization for a coral image predicted as infected with white band disease. The heatmap highlights the white band at the edge of the coral branches, which corresponds to necrotic tissue caused by the disease. The Grad-CAM shows that the proposed model correctly attends to these pathological features, effectively identifying the diseased tissue. This visualization demonstrates that the proposed framework learns biologically meaningful features for white band disease detection. By highlighting white band while ignoring healthy tissue and background, the framework provides an interpretable explanation of its decision-making, aligning well with expert understanding of disease pathology.

5. CONCLUSION

This paper has presented a novel approach for coral reef type classification as well as white band disease detection by using DL architectures. The proposed framework employed a two-stage ViT model for both coral reef classification and disease detection, along with XAI to enhance interpretability. Coral images underwent preprocessing to improve image quality. Performance evaluation was conducted using standard metrics, including accuracy, precision, recall, and F1-score. Experimental results demonstrated that the proposed model achieved superior performance compared to existing methods in both coral reef classification and white band disease detection. Grad-CAM visualizations successfully identified and highlighted regions associated with disease. It provides insight into the model's decision-making process and enhancing practical applicability. Future work will focus on exploring hybrid models that integrate multiple deep learning architectures to further improve classification and disease detection accuracy. Expansion of datasets and incorporation of multimodal data, such as environmental parameters, have the potential to provide a more comprehensive understanding of coral reef health and contribute to more effective conservation strategies.

REFERENCES

1. N. J. Macknight, K.Cobeleigh, D.Lasseigne. Microbial dysbiosis reflects disease resistance in diverse coral species. Commun Biol, 4,2021.
2. J.Justin, S.Maharjan, W.Li,E.Linstead, and S.P.Tiwar. et al.

A generalized machine learning model for long-term coral reef monitoring in the red sea. Heliyon, 10(18),2024

3. K.Sreekumar, C.Arvind,P.Anusha, K.Kishore, S.Chandragandhi, and K.Srihari. Classification of coral reefs in marine environments using deep encoder-decoder mechanism. J. of Survey in Fisheries Sci., 10(4S),2023,364-370
4. S. Jamil, M. Rahman, and A. Haider. Bag of Features (BoF) Based Deep Learning Framework for Bleached Corals Detection. Big Data Cogn. Comput. 5, 53,2021.
5. Fawad, I.Ahmad, A.Ullah, and W. Choi. Machine learning framework for precise localization of bleached corals using bag-of-hybrid visual feature classification. Sci. Rep., 13, 2023
6. M.H.Ibrahim and M.M.Sathik. A novel approach for coral reef disease detection and classification using deep learning techniques. 18(6),2021, 10029-10042
7. A. Chowdhury, M. Jahan, S. Kaisar, M.E. Khoda, S.M.A.K. Rajin, R. Naha, R. Coral Reef Surveillance with Machine Learning: A review of datasets, techniques, and challenges. Electronics, 13, 2024
8. A.Aldhahri, E.Saif, H.Ali,M.Alsayed and F.Alshareef. Harnessing computer vision and deep learning to monitor coral reef health. Engg. Tech. & Applied Sci. Res., 15(4),2025,24523-24531
9. S. Wang, N.-L. Chen, Y.-D.Song, T.-T.Wang, J. Wen, T.-Q. Guo, H.-J. Zhang, L. Mo, H.-R.Ma, and L. Xiang, L. ML-Net: A multi-local perception network for healthy and bleached coral image classification. J. Mar. Sci. Eng. 12, 2024.
10. S.N. Karthik, M. Hariharasudhan, M.A. Devi. An Investigation on Coral Reef Classification Using Machine Learning Algorithms. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. ICICC 2024. Lecture Notes in Networks and Systems, vol 1039.2025 Springer
11. G.A.Trudeau, K.Lowell, and J.A.Dijkstra. Coral reef detection using ICESat-2 and machine learning. Ecological Informatics, 87,2025.
12. M.S.Ogidi, and M.Sah. Binary classification of coral reef using deep learning for enhanced monitoring. 2025 9th Int. Symposium on Innovative Approaches in Smart Technologies.2025.
13. R.J.Firdous and S.Sabena. A novel approach to coral species classification using deep learning and unsupervised feature extraction. Journal of Spatial Science,2024,1-28.
14. Tameswar, K., Suddul, G., & Dookhitram, K. (2022). A hybrid deep learning approach with genetic and coral reefs metaheuristics for enhanced defect detection in software. International Journal of Information Management in Data Insights, 2