JOIV : Int. J. Inform. Visualization, 8(2) - May 2024 700-709

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

# Multi-Head Attention in Residual Networks to Improve Coral Reef Structure Classification

Eka Qadri Nuranti [a,*], Naili Suri Intizhami [a], Muhammad Irpan Sejati Tassakka [b], Intan Sari Areni [c], Osama Iyad Al Ghozy [a], Muhammad Rivaldi Jefri [a]

[a] *Department of Computer Science, Institut Teknologi Bacharuddin Jusuf Habibie, Bacukiki, Parepare, South Sulawesi, Indonesia*
[b] *Marine Engineering, Politeknik Perikanan dan Kelautan Bone, East Tanete Riattang, Bone, South Sulawesi, Indonesia*
[c] *Department of Electrical Engineering, Universitas Hasanuddin, Bontomarannu, Gowa, South Sulawesi, Indonesia*

*Corresponding author: *eka.qadri@ith.ac.id*

***Abstract*—Residual Networks (ResNet) mark a crucial advancement in convolutional neural network architecture, effectively tackling challenges like vanishing gradients for improved pattern detection in various image classification tasks. This study introduces a novel adaptation of the ResNet50 architecture that integrates a multi-head attention mechanism (MHA), coined MHA-ResNet50, for discerning coral reef structures within images. Strategic modifications are applied to the input of each stage, leading to the development of an MHA block, which is augmented by separable convolution. The deliberate inclusion of the MHA block at various stages in identity-block Resnet50, in adherence to multiscale gate principles, precedes its traversal through fully connected layers. Furthermore, we implemented the Stratified K-fold concept to ensure that each fold has a comparable proportion of each class. We successfully assessed the efficacy of the MHA-Resnet50 model in several MHA-block placement scenarios and saw improvements in the accuracy of coral reef structure predictions. The most optimal results were achieved by incorporating four attention blocks (MHA-ResNet50-4), yielding an accuracy rate of 85.23% in recognition of coral structure images, comprising a mere 409 images. This model showcases adaptability to small datasets while delivering commendable performance. The ResNet50 architecture undergoes enhancement in our proposed model by integrating multi-head attention, separable convolution, and multiscale gate principles. The MHA-ResNet50 model substantially advances accurately predicting coral reef structures, demonstrating adaptability to limited datasets. Future lines of this research involve digging deeper into the model design and using more significant amounts and classes of data to strengthen a more comprehensive range of generalizations.**

***Keywords*— Residual networks; convolutional neural network; attention mechanism; coral reef classification.**

*Manuscript received 30 Nov. 2023; revised 6 Jan. 2024; accepted 10 Feb. 2024. Date of publication 31 May 2024.*


## I. INTRODUCTION

Coral reefs are diverse marine ecosystems with great biological value [1]. Coral reefs provide food and habitat for numerous creatures [2]–[4]. However, global climate change, industrialization, population growth [5], and other factors have caused coral reef ecosystem devastation [1], [6]. Significant advancements have been made in image processing for underwater life studies in recent years [1], [7]. Examining the taxonomy of marine biota species [8], [9] or simply identifying the health of the coral reef [10] has served a crucial purpose in facilitating the analysis of interspecies distinctions and the conservation of endangered species. Using deep learning methodologies for processing purposes represents a preliminary phase in advancing maritime intelligent systems, specifically in artificial intelligence (AI) applications within undersea ecosystems.

The field of coral reef classification is now seeing notable advancements, offering possible opportunities for preserving and protecting these ecosystems [9]–[11]. The application of image classification techniques may provide advantageous outcomes in monitoring coral reefs and identifying areas requiring conservation efforts [8], [9], [12]. Environmental monitoring is also the first step in mapping coral reefs [13] and annotating them automatically [14], [15]. Additional studies aim to assess coral reefs' health [10] and resilience in the face of environmental and climate change stress by systematically monitoring observable changes in various coral reef species [5]. One of the implementation steps is to carry out feature extraction analysis on the color and texture of coral images [17]–[19]. Identifying the structure of coral

reefs is another challenging but equally fascinating task. Each species of coral reef may be identified structurally by having a branched physical structure [12], just like the *Acropora Cervicornis* and *Millepora Alcicornis* Varieties.

Image classification often uses Convolutional Neural Networks (CNNs). CNN hierarchically extracts image features using convolutional layers. Several notable architectures based on CNN model are AlexNet [20], VGG (Visual Geometry Group) [21], and ResNet (Residual Network) [22], with ResNet addressing the vanishing gradient problem through skip connections.

This study presents a model that utilizes computer vision and deep learning techniques to classify coral reefs, focusing on task coral structure. Since researchers introduced the Attention Mechanism approach [23], deep learning models have shown increasingly enhanced results in handling natural language processing problems with textual data [24]. In recent times, researchers have been directing their efforts toward extending this approach to image data [25]–[29], focusing on enhancing the performance of deep learning models in this context.

The main contribution of this paper is improved coral image classification predictions by adding attention mechanisms to the residual network architecture, improving performance accuracy. Previous research [12] has demonstrated using several Resnet Architectures, including Resnet50 and Resnet152. The findings show that Resnet50 is the most accurate in classifying coral structure datasets. This study will employ the Resnet50 model [12] as a baseline model, recognized as the most effective model for analyzing coral structure photos. Furthermore, we will experiment with placing the multi-head attention block (MHA Block) on every stage in the Resnet50 architecture. Additionally, we will evaluate and compare the performance of the proposed model with existing coral image classification approaches to demonstrate its effectiveness and superiority in accurately identifying coral species.

This article has four distinct sections. The first section encompasses an introductory segment that provides contextual information and comprehensively examines the relevant works. The subsequent section will clarify the proposed approach to improving the accuracy of coral reef image classification by utilizing ResNet Architecture and Multi Head Attention. The next part will discuss the experiment's outcomes and the following analysis. Lastly, the final section will dive into the implications and potential avenues for further study.

## II. The Material and Method

We aim to consider the success of research that successfully applies multi-head attention to CNN architecture for recognizing COVID-19 [27] and Human Activity Recognition [29]. Our proposed model improves the ResNet50 architecture by including multi-head attention, separable convolution, and multi-scale gate principles. The purpose model is called MHA-Resnet50.

### *A. Residual Neural Network (ResNet)*

The Residual Neural Network (ResNet) [30] is a specific Convolutional Neural Network (CNN) design incorporating skip connections to learn residual functions concerning the layer inputs. Increasing model training depth with deep neural networks can lead to overfitting and gradient vanishing. The idea of skipping connections can overcome these limitations. Li's study [31] Implementing skip connections effectively demonstrated the visualization of the outcomes of the loss function in a neural network. Skip connections deliver a perceptible enhancement of the smoothness of the loss surface. Furthermore, skip connections will also impact the velocity of the network.

The original architecture of ResNet is ResNet34 [22], which contains numerous convolutional layers. In recent years, researchers have assembled several ResNet architectures, including ResNet50. As shown in Fig. 1, ResNet50 has two primary blocks: identity and convolution. Both blocks implement a skip connection; input passes through two main and short paths. A short path is a direct path from input to output without transformation. These two pathways are added together. The function of the short path in the identity block is to keep the original information from the input. This solves the backpropagation gradient vanishing problem. On the short path in the convolutional block, a convolution layer is applied to change the input's dimensions or resolution to match the primary path's output. After that, the two are also added together. The short path makes the input dimensions match the primary path's production. ResNet uses skip connections to train intensive networks with stable gradient flow.

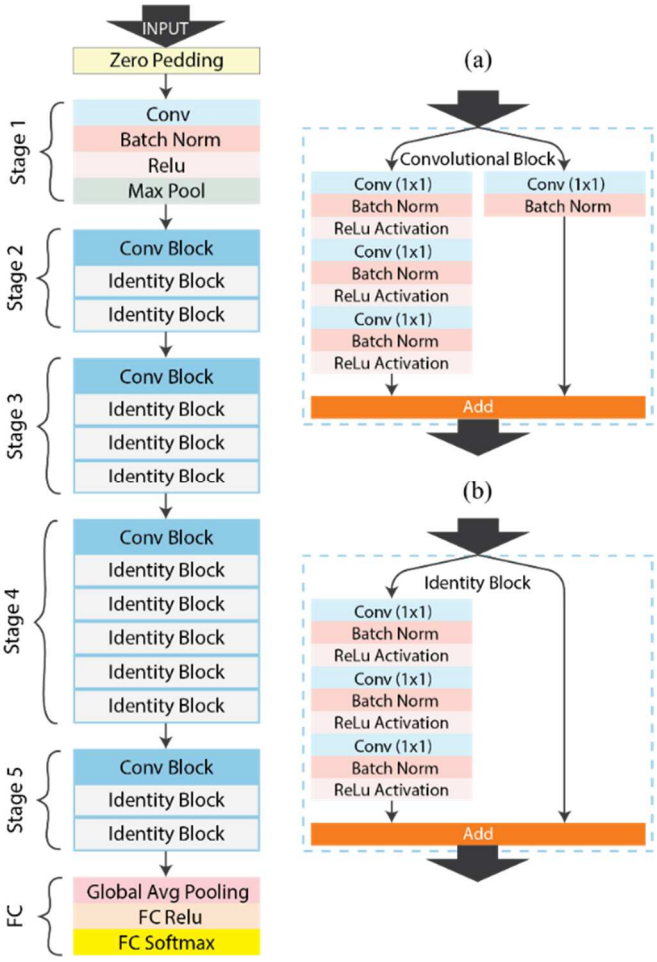


Fig. 1 Resnet50 Architecture,
(a) Convolutional Block or Conv Block, and (b) Identity Block

We set the fully connected layer in Resnet50 by replacing 1000 neurons with 512 neurons and using ReLu activation according to the settings in the previous paper [12].

### *B. Depth-Separable Convolutional*

Depth-separable convolution is a neural network technique that breaks down the standard convolution into two separate operations: depth wise convolution and pointwise convolution, which can be seen in Fig. 2.

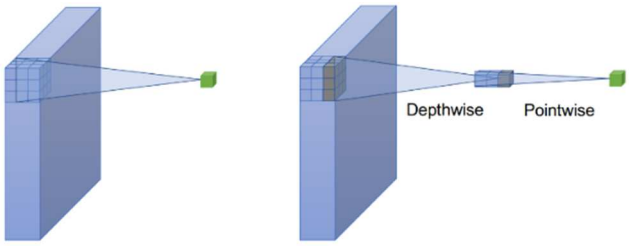


Fig. 2 Standard convolution and separable convolution [33]

During the depthwise, a separate convolution operation is performed on each input channel using a 1x1x[input_channels] filter. This helps to decrease the computational workload. Subsequently, pointwise convolution combines the output channels of the 1x1 convolution kernel into a condensed form. This separation reduces the number of parameters and enables the model to capture map features more efficiently [32].

### *C. Multi-Head Attention*

The attention mechanism functions similarly to human vision, prioritizing the focal point of interest rather than processing the complete information in the picture [23]. Hence, the primary objective of the attention mechanism is to dynamically choose essential features in the image that substantially impact the prediction outputs [27]. Within its implementation, there exist three inputs within a feature map that contain the same information, specifically Query ($Q$), Key ($K$), and Value ($V$). These representations are referred to as key-value queries. The score for each element in the query matrix ($S_i$) is computed by performing a dot product between the query and the transpose of the matrix key ($K_i^T$). The equations involve using the variable $i$, which represents the attention index. The scores obtained are normalized by dividing them by the square root of the vector's key dimension ($d_k$). Subsequently, these normalized scores are processed through a softmax function to calculate the attention weights; we can see Eq. 1. Ultimately, the attention weights are employed to compute the weighted summation of the values, resulting in the output of the attention mechanism as Eq. 2.

$$S_i = \text{softmax}\left(\frac{Q \times K_i^T}{\sqrt{d_k}}\right) \quad (1)$$

$$Attention(Q, K, V) = S_i \times V_i \quad (2)$$

The multi-head attention (MHA) has numerous attention heads that calculate attention weights for different input segments. In the multi-head attention mechanism, every head (Eq. 3) possesses projection matrices represented as $W_i Q$, $W_i K$, and $W_i V$. The equations involve the variable h, which indicates the number of attention heads, the visual representation in **Error! Reference source not found.**.

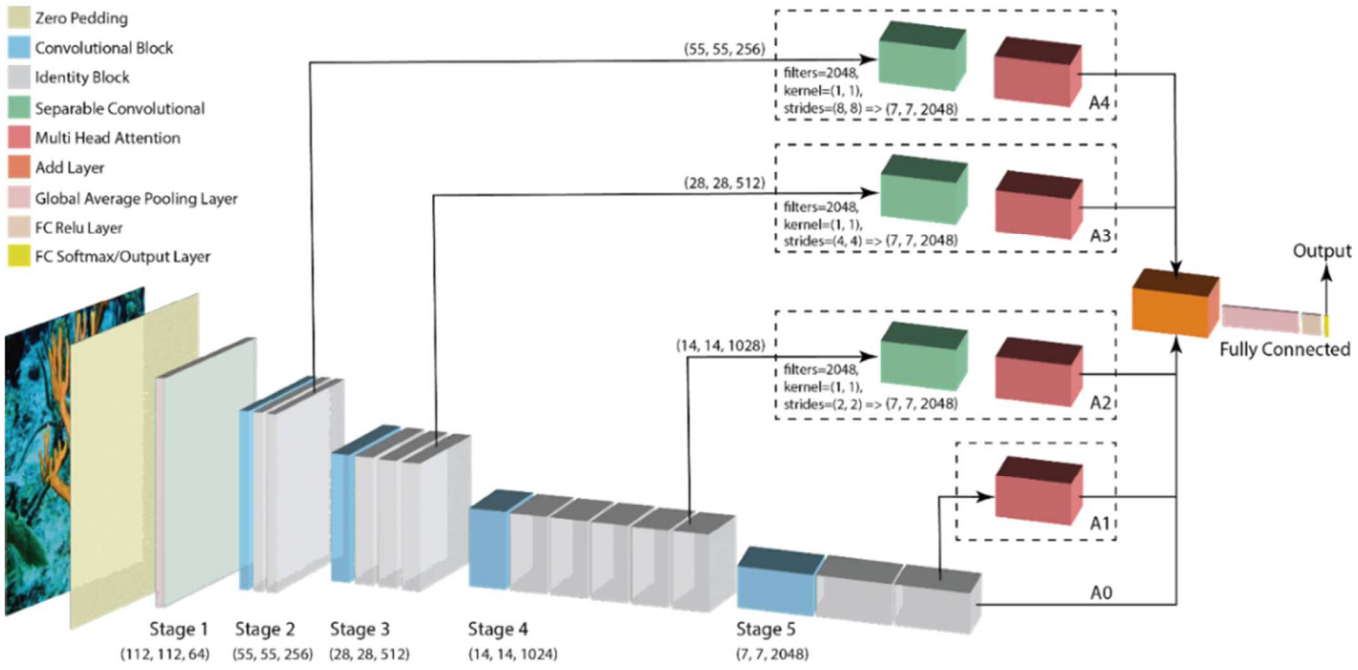


Fig. 3 Structure of Purpose Model: Multi Head Attention Resnet50

The MHA equation (Eq. 4) allows the model to gather information on the connections between the available data. Equation 4 employs a concatenation operation, denoted by $Concat$, and utilizes the weight matrix $W^O$ to perform the ultimate linear transformation. This MHA will ensure that no information is lost during the process.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

$$MHA = Concat(head_1, \dots, head_h) \times W^O \quad (4)$$

*D. Proposed Method*

Fig. 3 depicts the structural design of our model and the output shape of every stage in TABLE I, highlighting the incorporation of the multi-head attention Resnet50. In Resnet50, we add a Multi-head Attention (MHA) block that is smoothly integrated after each stage's last identity block. To ensure smooth integration at every level, it is crucial to make necessary adjustments to the input size. Nevertheless, we intend to avoid the inclusion of complex calculations. Thus, we choose a practical and effective strategy by utilizing separable convolution as the most appropriate method for this stage.

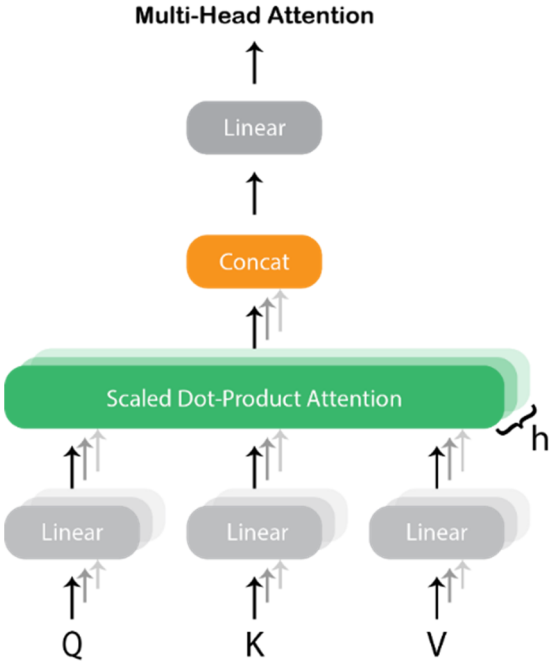


Fig. 4 Multi-Head Attention

TABLE I
LAYER OUTPUT OF MULTI HEAD ATTENTION IN RESNET50

| Block | Layer (type) | Output Shape |
|---|---|---|
| | Input Layer | (None, 224, 224, 3) |
| | Zero Padding | (None, 230, 230, 3) |
| Stage 1 | Conv2D | (None, 112, 112, 64) |
| | Batch Norm | (None, 112, 112, 64) |
| | Relu | (None, 112, 112, 64) |
| | Max Pool | (None, 56, 56, 64) |
| Stage 2 | Conv Block | (None, 56, 56, 256) |
| | Identity Block x2 | (None, 56, 56, 256) |
| Stage 3 | Conv Block | (None, 28, 28, 512) |
| | Identity Block x3 | (None, 28, 28, 512) |
| Stage 4 | Conv Block | (None, 14, 14, 1024) |
| | Identity Block x5 | (None, 14, 14, 1024) |
| Stage 5 | Conv Block | (None, 7, 7, 2048) |
| | Identity Block x2 | (None, 7, 7, 2048) |
| A1 | MHA A1 | (None, 7, 7, 2048) |
| A2 | Separable Conv2D | (None, 7, 7, 2048) |
| | MHA A2 | (None, 7, 7, 2048) |
| A3 | Separable Conv2D | (None, 7, 7, 2048) |
| | MHA A3 | (None, 7, 7, 2048) |
| A4 | Separable Conv2D | (None, 7, 7, 2048) |
| | MHA A4 | (None, 7, 7, 2048) |
| FC | Add | (None, 7, 7, 2048) |
| | Global Avg Pooling | (None, 2048) |
| | Dense (FC Relu) | (None, 512) |
| | Dense (Output) | (None, 14) |

Multiscale gate refers to analyzing the performance of MHA at several stages, considering different attention. Various attention blocks can be utilized to implement MHA-Resnet50. MHA-Resnet50-1 means structure including 'A0' and 'A1' in Table 1, MHA-Resnet50-2 means structure including 'A0', 'A1' and 'A2', MHA-Resnet50-3 means structure including 'A0', 'A1', 'A2' and 'A3 ', and MHA-Resnet50-4 means the structure includes 'A0', 'A1', 'A2', 'A3' and 'A4' in TABLE I. In outline, the analysis scenario refers to TABLE II. The purpose of each MHA block is to carefully modify a feature map of many sizes by incorporating the Add layer, hence efficiently practicing the skip connection concept. The strategic implementation aims to maintain both the simplicity of the model and the complex interconnection of information. The previous study [12] we assessed the StructureRSMAS dataset using five deep learning models (Inception, ResNet-50, ResNet-152, DenseNet-121, and DenseNet-161). The most favorable outcomes were achieved by employing ResNet-50 with a batch size of 32 and 300 epochs. We utilized the identical model setup, with a batch size of 32, and implemented early stopping as part of the model training. The selection of epochs is contingent upon the model's ability to learn. We utilized the accuracy matrix as the monitoring parameter for the early-stopping technique we deployed. The training procedure will cease when the accuracy shows minor improvement, enabling the model to reduce loss.

TABLE II
SCENARIO MODEL MULTI HEAD ATTENTION IN RESNET50

| No. | Model Name | Multi Scale |
|---|---|---|
| 1 | ResNet50 | Baseline (A0) |
| 2 | MHA-ResNet50-1 | A0+A1 |
| 3 | MHA-ResNet50-2 | A0+A1+A2 |
| 4 | MHA-ResNet50-3 | A0+A1+A2+A3 |
| 5 | MHA-ResNet50-4 | A0+A1+A2+A3+A4 |

## *E. Dataset*

The dataset employed in this study is the StructureRSMAS dataset [12], which comprises coral structures. The dataset was obtained using different cameras and varied conditions. This situation is commendable and is anticipated to align with the authentic portrayal. A total of 409 coral photos have been captured, providing explicit visual representations of the structural characteristics shown by 14 distinct coral classes. Image distribution can be seen in TABLE III. Unbalanced data includes the availability of class images. The most significant number of available images is in the ACER class, with 44 photos; the smallest is in the DANT class, namely 16 images. In StructureRSMAS, we can see one of each coral class listed in Fig. 5.

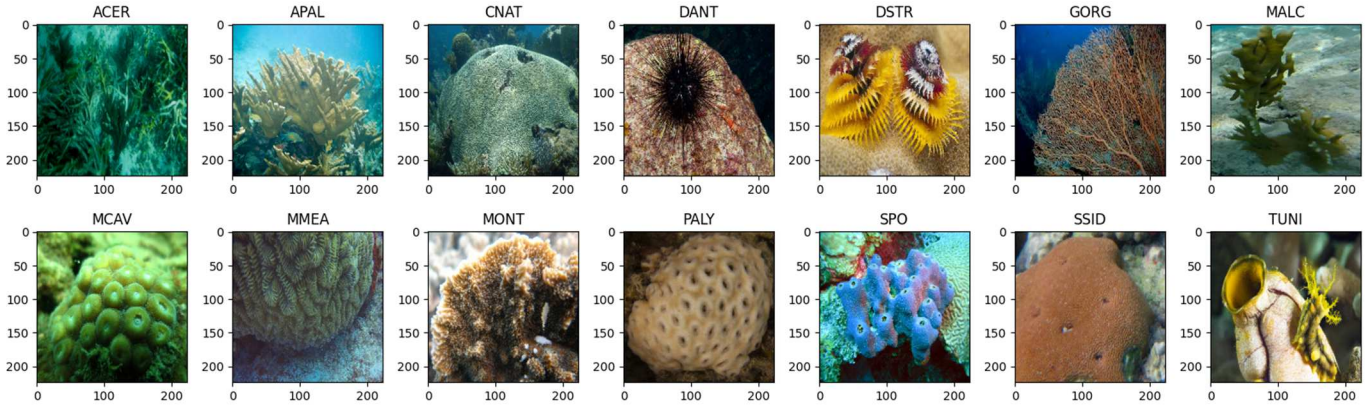


Fig. 5 One image of each coral reef structure class

TABLE III
CLASS DISTRIBUTION ON THE STRUCTURERSMAS DATASET

| Classes | #imgs |
|---|---|
| Acropora Cervicornis (ACER) | 44 |
| Acropora Palmata (APAL) | 41 |
| Colpophyllia Natans (CNAT) | 34 |
| Diadema Antillarum (DANT) | 20 |
| Diploria Strigosa (DSTR) | 16 |
| Gorgonians (GORG) | 18 |
| Millepora Alcicornis (MALC) | 33 |
| Montastraea Cavernosa (MCAV) | 38 |
| Meandrina Meandrites (MMEA) | 30 |
| Montipora spp. (MONT) | 21 |
| Palythoas Palythoa (PALY) | 32 |
| Sponge Fungus (SPO) | 23 |
| Siderastrea Siderea (SSID) | 36 |
| Tunicates (TUNI) | 23 |

## *F. Stratified K-fold*

To begin the pre-processing data, the first stage consisted of dividing the dataset into training and testing sets, with 80% of the data allocated for training and 20% for testing for each class. The training dataset was treated to the Stratified K-fold methodology [34], precisely a 5-fold approach. This approach enabled the division of the training data into separate subsets for training and validation, ensuring a fair distribution across different categories. Fig. 6 visually represents the final data distribution, offering a thorough perspective of the balanced representation.

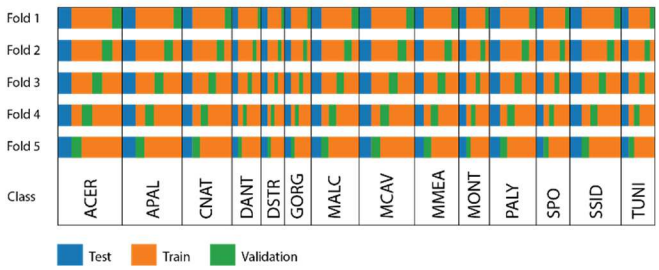


Fig. 6 Visualization of train, validation, and test datasets using stratified 5-fold

## III. Results and Discussion

In this section, we discuss and evaluate the results of the proposed method. The evaluation metrics employed include accuracy (acc), precision (pre), recall (rec), and f1-score (f1). These matrices offer a comprehensive perspective on the performance of the proposed approach, allowing for a holistic evaluation of the accuracy and precision of the model.

To begin with, we present the performance outcomes of the model using ResNet and the proposed method MHA-ResnNet50. Our comparison is limited to the ResNet50 model because the prior study, [12] analyzed five distinct models. Next, we examine the performance of the most well-trained and validated models for each epoch across the data folds. Additionally, we present the model's performance on 14 different coral structure classes using a confusion matrix and show precision, recall, and f1-score evaluations for each class. Finally, we evaluate instances where the proposed model's predictions differ from the actual coral classes.

TABLE IV compares the original ResNet and the multi-scale approach in MHA-Resnet50 based on the model scenarios provided in TABLE II. After analyzing the performance metrics of ResNet and the four multi-scale MHA-Resnet50 versions, it is evident that including multi-head attention has a beneficial effect on ResNet50's predictions. At every stage of the MHA block, it significantly improves ResNet50's accuracy metrics, with enhancements ranging from 1.14 to 5.69. In addition, the MHA blocks at each stage have a significant function, resulting in a maximum accuracy of 85.23% for MHA-Resnet50-4.

TABLE IV
Model Performance Comparison

| No | Model Name | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| 1 | ResNet50 | 79.54 | 82.86 | 79.54 | 79.37 |
| 2 | MHA-ResNet50-1 | 80.68 | 83.57 | 80.68 | 80.19 |
| 3 | MHA-ResNet50-2 | 81.82 | 82.43 | 81.82 | 81.63 |
| 4 | MHA-ResNet50-3 | 84.09 | 84.77 | 84.09 | 83.62 |
| 5 | MHA-ResNet50-4 | 85.23 | 86.23 | 85.23 | 85.01 |

An in-depth analysis of the chart shown in Fig. 7 provides a detailed evaluation of the accuracy performance demonstrated by each model in different data folds. Every model demonstrates excellent performance, keeping a stable quality of accuracy through the folds, except for a notable decline in the accuracy of MHA-Resnet50-2 during the third fold. Considering the model that achieves the maximum level of accuracy, specifically MHA-Resnet50-4, its accuracy continually increases in each fold until it achieves its peak performance. On the other hand, various models exhibit different trends of improving and decreasing accuracy. This perceptive remark highlights the beneficial influence of incorporating Blocks A0, A1, A2, A3, and A4 on the overall effectiveness of the model.

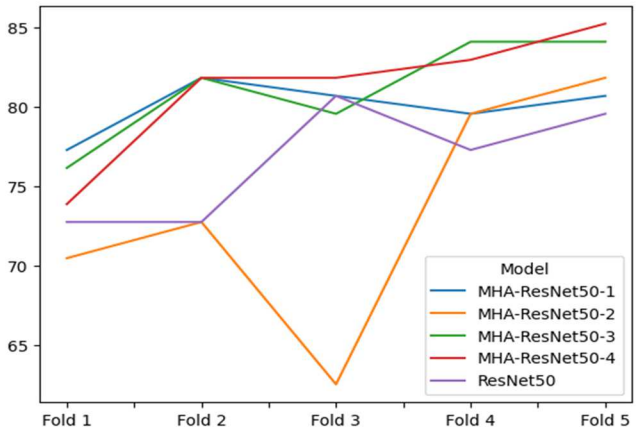


Fig. 7 Graph of stratified 5-fold accuracy against the models

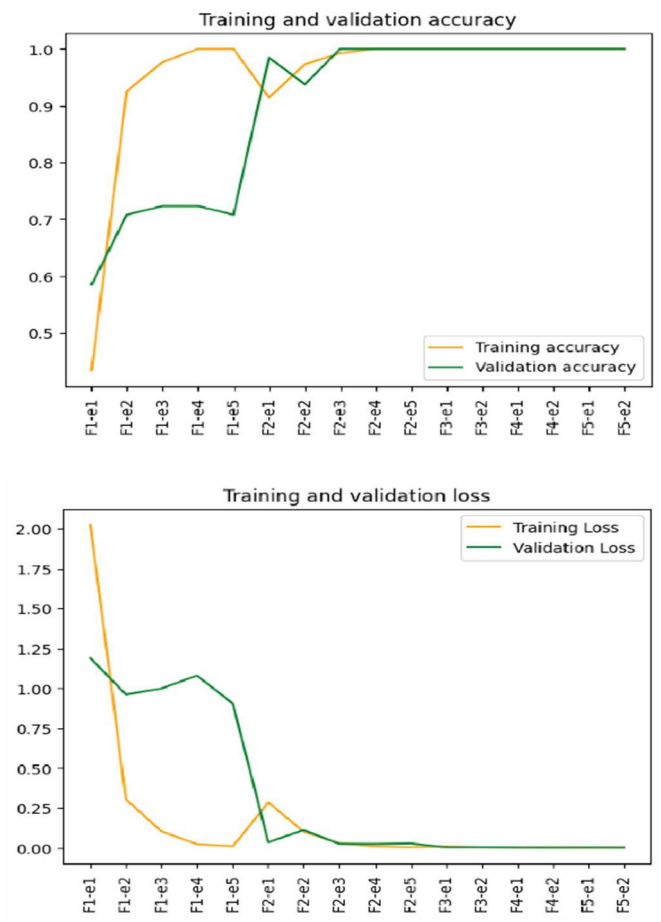


Fig. 8 Graph of train and validation at each fold and epoch for the MHA-ResNet50-4; the top image is for the accuracy value, and the bottom matrix is for the loss value

Fig. 8 illustrates the loss and accuracy graphs seen during the training process. It shows a distinct pattern in the number of epochs for each data fold. This pattern comes from the utilization of early stopping, a method that enables to stop its training early if there is no substantial enhancement in accuracy on the validation set or if the accuracy value on the validation set fails to exceed the value in the previous epoch. For instance, during the initial stage, specifically from epoch 4 (F1-e4) to epoch 5 (F1-e5), there is a noticeable decrease in accuracy when evaluating the validation data. The early stopping condition is activated, causing the training process to stop at epoch 5. Early stopping is a preventative strategy to prevent overfitting, which occurs when the model overly adjusts to the training data but fails to perform well on unseen information. Hence, the variation in epoch numbers observed in each fold in Fig. 8 can be assigned to the sensitivity of early stopping to changes in the model's performance on the validation set, supporting a more efficient model and avoiding excessive adaptation.

TABLE V shows a thorough evaluation of the prediction accuracy of the MHA-ResNet50-4 model when applied to the testing dataset. The matrix performance of all coral classes is evaluated using weighted average scores, which provide an overall perspective on the model's performance. The relevant classes' precision, recall, and F1-Score values are 0.87, 0.85, and 0.85. The precision metrics reflect the model's ability to correctly recognize instances of each class, with significant cases of perfect precision (1.00) for classes APAL, DANT, DSTR, SPO, and TUNI. The precision scores show high accuracy in the model's positive predictions for these classes. Recall values indicate the model's ability to identify all relevant examples of each class accurately. Notably, the classes DANT, DSTR, GORG, MALC, PALY, SPO, and TUNI achieved a perfect recall score of 1.00, meaning that the model accurately identified all instances belonging to these classes. The F1-Score, determined as the harmonic mean of precision and recall, offers a well-balanced evaluation of the model's performance. Classes DANT, DSTR, GORG, MALC, PALY, SPO, and TUNI tend to demonstrate exceptional F1-Scores, suggesting a well-balanced combination of precision and recall for these classes. The highest F1-Score values are achieved for three coral classes: DANT, DSTR, and TUNI.

TABLE V
PERFORMANCE EVALUATION OF PREDICTION CLASSES

| Classes | Precision | Recall | F1-Score |
|---|---|---|---|
| ACER | 0.89 | 0.73 | 0.80 |
| APAL | 1.00 | 0.75 | 0.86 |
| CNAT | 0.86 | 0.86 | 0.86 |
| DANT | 1.00 | 1.00 | 1.00 |
| DSTR | 1.00 | 1.00 | 1.00 |
| GORG | 0.75 | 0.75 | 0.75 |
| MALC | 0.71 | 1.00 | 0.83 |
| MCAV | 0.62 | 0.83 | 0.71 |
| MMEA | 0.83 | 0.83 | 0.83 |
| MONT | 0.60 | 0.75 | 0.67 |
| PALY | 0.86 | 0.86 | 0.86 |
| SPO | 1.00 | 0.83 | 0.91 |
| SSID | 0.88 | 1.00 | 0.93 |
| TUNI | 1.00 | 1.00 | 1.00 |
| Accuracy | | | 0.85 |
| Macro Avg. | 0.86 | 0.87 | 0.86 |
| Weight Avg. | 0.87 | 0.85 | 0.85 |

The model consistently demonstrates high predictive capability across various coral architectures, achieving commendable precision, recall, and F1-Score scores. The balanced macro and weighted averages further support the model's success in resolving class imbalances and delivering accurate forecasts throughout the dataset. Furthermore, the model's strong performance in many evaluation criteria highlights its dependability and appropriateness for classifying coral reef structures in real-world scenarios.

Figure 9 compares the confusion matrices produced by two different models: MHA-Resnet50-4 and ResNet50. These matrices are depicted in Fig.9(2) and Fig.9(1) respectively. The analysis goes beyond simple numerical comparisons and explores the qualitative components of prediction performance. After examining the predictions made on 88 test photos, it was discovered that ResNet50 made 18 incorrect predictions. However, the MHA-Resnet50-4 model, albeit showing improved accuracy, still exhibits 13 mistakes. Although the two models exhibited comparable mistakes, it is interesting that MHA-Resnet50-4 outperformed in many cases by successfully identifying specific data points. This comparative analysis provides insight into the subtle variations in performance between the two models, as evidenced in Table 6.

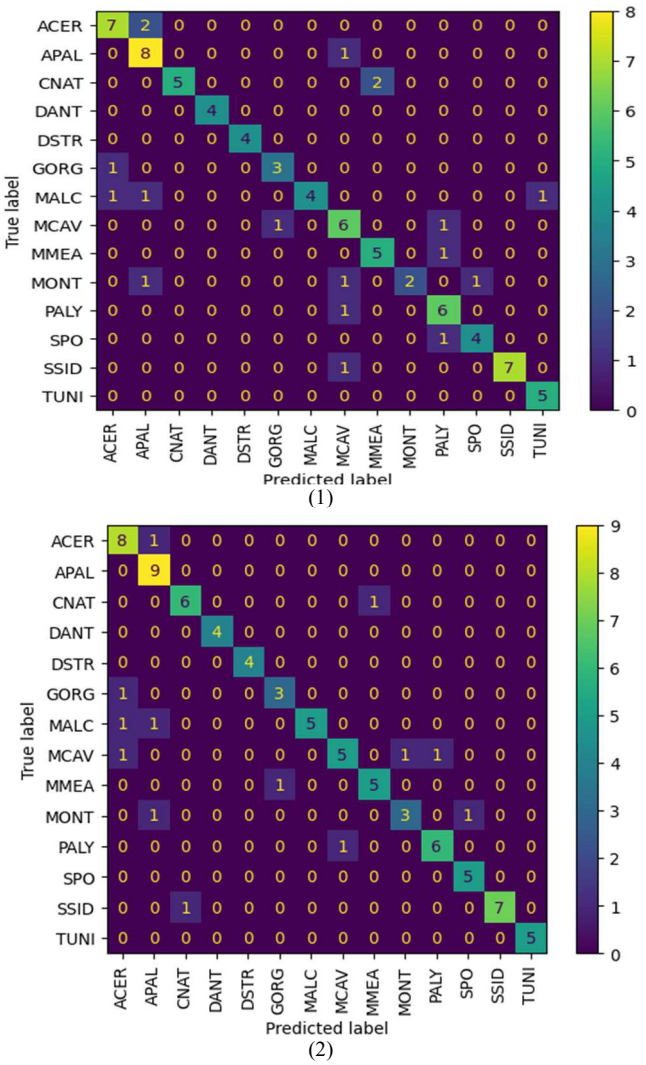


Fig. 9 Confusion Matrix of (1) Resnet50 and (2) Purpose Method MHA-Resnet50-4

The changes in prediction accuracy are presented in Table 6 to elucidate the reasons for misclassification. The table provides a comprehensive overview of the prediction results for ResNet50 and MHA-Resnet50-4 models. Precise categorizations are indicated by bold, but incorrect categorizations are not. This deliberate distinction enables a comprehensive examination of the models' performance, enhancing our comprehension of their relative effectiveness. Upon close examination, it is evident that MHA-Resnet50-4 has precise classification capabilities by accurately identifying and categorizing five image files that the original model previously misclassified. Nevertheless, MHA-Resnet50-4 exhibits misclassification in two image files, indicating challenges in accurately forecasting results.

TABLE VI
COMPARISON OF PREDICTION RESULTS FROM RESNET50 AND MHA-RESNET50-4

| Class | File Name | Prediction Result | |
|---|---|---|---|
| | | ResNet50 | MHA-ResNet50-4 |
| ACER | acer35 | APAL | APAL |
| | acer38 | APAL | **ACER** |
| APAL | apal15 | MCAV | **APAL** |
| CNAT | cnat13 | MMEA | **CNAT** |
| | cnat30 | MMEA | MMEA |
| GORG | gorg12 | ACER | ACER |
| | malc02 | TUNI | APAL |
| MALC | malc13 | ACER | ACER |
| | malc15 | ACER | **MALC** |
| | mcav04 | PALY | ACER |
| MCAV | mcav30 | **MCAV** | MONT |
| | mcav38 | PALY | PALY |
| MMEA | mmea18 | MCAV | GORG |
| | mont15 | MCAV | **MONT** |
| MONT | mont18 | MCAV | SPO |
| | mont19 | APAL | APAL |
| PALY | paly15 | MCAV | MCAV |
| SPO | spo07 | PALY | **SPO** |
| SSID | ssid04 | **SSID** | CNAT |
| | ssid30 | MCAV | **SSID** |

Table 6 comprehensively assesses prediction errors by providing specific information on misclassified files. These errors provide insight into the subtle variations in performance between the ResNet50 and MHA-ResNet50-4 versions. An additional understanding of these inconsistencies can be obtained from Fig. 10, which graphically illustrates the exact occurrences of misclassification. Fig. 10 (1) demonstrates five instances of prediction mistakes in which ResNet50 failed, whereas MHA-ResNet50-4 excelled, highlighting the latter's superior predictive ability. In contrast, Fig. 10(2) demonstrates two specific cases where MHA-ResNet50-4 encountered difficulties while ResNet50 succeeded, illustrating the intricacies of prediction in particular situations. By comparing Figure 10 with the coral class characteristics illustrated in Figure 5, we obtain visual observations of the prediction discrepancies among various classes. Specifically, Fig.10 (1) highlights the constraints of ResNet50 in precisely categorizing finger-shaped coral reef formations, such as the ACER, APAL, and MALC categories. This emphasizes the efficacy of MHA-ResNet50-4 in overcoming these deficiencies.

Furthermore, ResNet50 encounters difficulties when processing photos with complex textures instead of clear and well-defined structures, as observed in the MONT, SPO, and SSID coral classes. Although the suggested approach enhances predictions in specific categories, such as ACER and APAL, it does not correct inaccuracies in the SSID category, as indicated in Figure 10 (2). This highlights the need for additional research to enhance model performance accuracy across many coral reef formations.

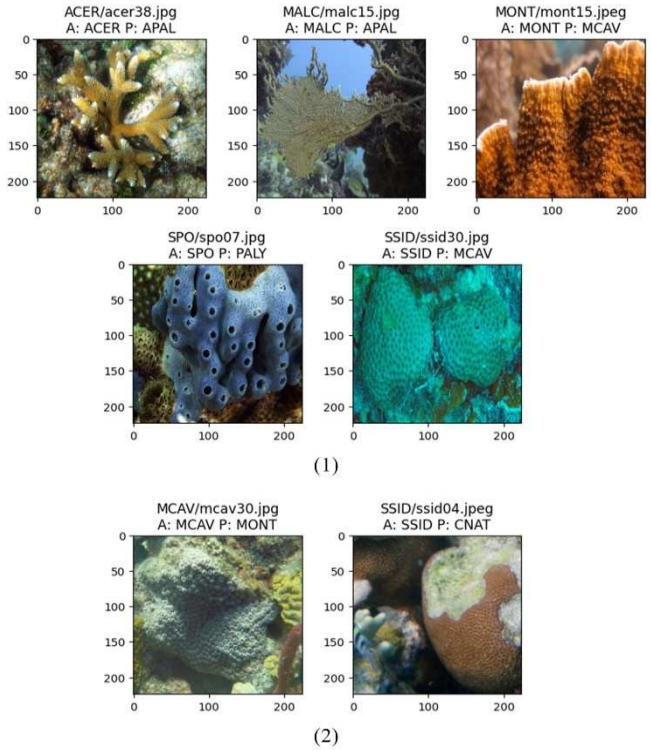


Fig. 10 (1) a prediction error occurs in ResNet50 but is predicted correctly by MHA-ResNet50-4 (2) a prediction error occurs in MHA-ResNet50-4 but is predicted correctly by ResNet50

## IV. Conclusions

This paper presents improvements to the ResNet50 Architecture to accurately identify and classify coral structure images. This work utilizes the StructureRSMAS dataset, which has 14 different coral classes. The dataset has a sample size of 409, enabling the model to achieve impressive performance despite its relatively small size. The proposed method shows superior performance compared to existing baseline methods. These findings confirm that adding Multi-Head Attention at each stage of the ResNet50 Architecture significantly improves the model's ability to recognize the coral structure, producing competitive results with an increase in accuracy of 5.69% compared to the ResNet50 model without modification. Further exploration of variations in model architecture can be carried out for future research; this is necessary as our current model has limitations in accurately identifying the complex characteristics of certain corals with similar structures.

## Acknowledgment

The authors thank the Hibah DRTPM 2023 (Grant No. 109/E5.5/PG.02.00.TL/2023) from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for the financial support. Additionally, the LPPM-PM of Institut Teknologi Bacharuddin Jusuf Habibie is appreciated for the comprehensive assistance. Their dedicated efforts have played a crucial role in successfully implementing this study.

## References

[1] T. N. T. Arsad, E. A. Awalludin, Z. Bachok, W. N. J. H. W. Yussof, and M. S. Hitam, "A review of coral reef classification study using deep learning approach," AIP Conference Proceedings, 2023, doi:10.1063/5.0110245.

[2] A. Triwibowo, "Strategi Pengelolaan Ekosistem Terumbu Karang di Wilayah Pesisir," Jurnal Kelautan dan Perikanan Terapan (JKPT), vol. 1, p. 61, Jan. 2023, doi: 10.15578/jkpt.v1i0.12048.

[3] J. Kleypas et al., "Designing a blueprint for coral reef survival," Biological Conservation, vol. 257, p. 109107, May 2021, doi:10.1016/j.biocon.2021.109107.

[4] Tri Aryono Hadi, Abrar Muhammad, Giyanto Giyanto, and Bayu Prayudha, *The Status of Indonesian Coral Reefs 2019*. Research Center for Oceanography, 2020. Accessed: Apr. 04, 2023. [Online]. Available: http://lipi.go.id/publikasi/status-of-indonesia-coral-reef/33174

[5] M. I. S. Tassakka, I. Alsita, S. Sahari, and A. K. Admaja, "Identification of Coral Reef Ecosystems Using the Coral Point Count Application with Excel Extensions (CPCe) on Wangiwangi Island, Wakatobi, Indonesia," 2023.

[6] M. C. Ladd and A. A. Shantz, "Trophic interactions in coral reef restoration: A review," Food Webs, vol. 24, p. e00149, Sep. 2020, doi:10.1016/j.fooweb.2020.e00149.

[7] A. King, S. M. Bhandarkar, and B. M. Hopkinson, "A Comparison of Deep Learning Methods for Semantic Segmentation of Coral Reef Survey Images," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2018, doi:10.1109/cvprw.2018.00188.

[8] T. Boone-Sifuentes et al., "Marine-tree: A Large-scale Marine Organisms Dataset for Hierarchical Image Classification," Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Oct. 2022, doi: 10.1145/3511808.3557634.

[9] A. Raphael, Z. Dubinsky, D. Iluz, and N. S. Netanyahu, "Neural Network Recognition of Marine Benthos and Corals," Diversity, vol. 12, no. 1, p. 29, Jan. 2020, doi: 10.3390/d12010029.

[10] JJ. Borbon, J. Javier, J. Llamado, E. Dadios, and R. K. Billones, "Coral Health Identification using Image Classification and Convolutional Neural Networks," 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Nov. 2021, doi: 10.1109/hnicem54116.2021.9731905.

[11] J. T. Ridge, P. C. Gray, A. E. Windle, and D. W. Johnston, "Deep learning for coastal resource conservation: automating detection of shellfish reefs," Remote Sensing in Ecology and Conservation, vol. 6, no. 4, pp. 431–440, Feb. 2020, doi: 10.1002/rse2.134.

[12] A. Gómez-Ríos, S. Tabik, J. Luengo, A. S. M. Shihavuddin, and F. Herrera, "Coral species identification with texture or structure images using a two-level classifier based on Convolutional Neural Networks," Knowledge-Based Systems, vol. 184, p. 104891, Nov. 2019, doi:10.1016/j.knosys.2019.104891.

[13] M. Gholoum, D. Bruce, and S. Alhazeem, "A new image classification approach for mapping coral density in State of Kuwait using high spatial resolution satellite images," International Journal of Remote Sensing, vol. 40, no. 12, pp. 4787–4816, Mar. 2019, doi:10.1080/01431161.2019.1574991.

[14] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2012, doi: 10.1109/cvpr.2012.6247798.

[15] L. Picek, A. Říha, and A. Zita, "Coral Reef annotation, localisation and pixel-wise classification using Mask R-CNN and Bag of Tricks".

[16] R. Ardiwidjaja, "Pelestarian Tinggalan Budaya Bawah Air: Pemanfaatan Kapal Karam sebagai Daya Tarik Wisata Selam," AMERTA, vol. 35, no. 2, p. 133, Dec. 2017, doi:10.24832/amt.v35i2.251.

[17] S. Jalali, P. J. Seekings, C. Tan, H. Z. W. Tan, J.-H. Lim, and E. A. Taylor, "Classification of marine organisms in underwater images using CQ-HMAX biologically inspired color approach," The 2013 International Joint Conference on Neural Networks (IJCNN), Aug. 2013, doi: 10.1109/ijcnn.2013.6707084.

[18] M. Soriano, S. Marcos, C. Saloma, M. Quibilan, and P. Alino, "Image classification of coral reef components from underwater color video," MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295), doi: 10.1109/oceans.2001.968254.

[19] A. Mahmood et al., "Coral classification with hybrid feature representations," 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, doi: 10.1109/icip.2016.7532411.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks,"

Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[21] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Nov. 2015, doi:10.1109/acpr.2015.7486599.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, doi: 10.1109/cvpr.2016.90.
[23] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[24] E. Q. Nuranti, E. Yulianti, and H. S. Husin, "Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents," Computers, vol. 11, no. 6, p. 88, May 2022, doi: 10.3390/computers11060088.
[25] Front. Plant Sci., vol. 11, p. 600854, Dec. 2020, doi:10.3389/fpls.2020.600854.
[26] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340–350, Oct. 2020, doi: 10.1016/j.neucom.2020.06.014.
[27] G. Hong, X. Chen, J. Chen, M. Zhang, Y. Ren, and X. Zhang, "A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19," Scientific Reports, vol. 11, no. 1, Sep. 2021, doi: 10.1038/s41598-021-97428-8.
[28] H. B. Khoirullah, N. Yudistira, and F. A. Bachtiar, "Facial Expression Recognition Using Convolutional Neural Network with Attention Module," JOIV : International Journal on Informatics Visualization, vol. 6, no. 4, p. 897, Dec. 2022, doi: 10.30630/joiv.6.4.963.
[29] T.-H. Tan, Y.-L. Chang, J.-R. Wu, Y.-F. Chen, and M. Alkhaleefah, "Convolutional Neural Network With Multihead Attention for Human Activity Recognition," IEEE Internet of Things Journal, vol. 11, no. 2, pp. 3032–3043, Jan. 2024, doi: 10.1109/jiot.2023.3294421.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. Accessed: Oct. 17, 2023. [Online]. Available: http://arxiv.org/abs/1512.03385
[31] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the Loss Landscape of Neural Nets".
[32] N. S. Inthizami et al., "Flood video segmentation on remotely sensed UAV using improved Efficient Neural Network," ICT Express, vol. 8, no. 3, pp. 347–351, Sep. 2022, doi: 10.1016/j.icte.2022.01.016.
[33] Y. Guo, Y. Li, L. Wang, and T. Rosing, "Depthwise Convolution Is All You Need for Learning Multiple Visual Domains," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8368–8375, Jul. 2019, doi: 10.1609/aaai.v33i01.33018368.
[34] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," Frontiers in Nanotechnology, vol. 4, Aug. 2022, doi:10.3389/fnano.2022.972421.