BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2709–2726
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss4pp2709-2726

# CLASSIFICATION OF SKELETAL MALOCCLUSION USING CONVENTIONAL NEURAL NETWORK (CNN) WITH VISION ATTENTION

**I Putu Ronny Eka Wicaksana**[1*], **Antoni Wibowo**[2], **Rojali**[3], **Azurah A Samah**[4], **Aspalilah Alias**[5]

*[1,2,3] Computer Science Department, BINUS Graduate Program–Master of Computer Science,*
*Bina Nusantara University*
*Jln. K. H. Syahdan No. 9, Jakarta Barat, 11480, Indonesia*

*[4]Faculty of Computing, Universiti Teknologi Malaysia*
*Johor Baru, 81310, Malaysia*

*[5]Department of Basic Science and Oral Biology, Faculty of Dentistry, Universiti Sains Islam Malaysia*
*Jln. Pandan Utama, Kuala Lumpur, 55100, Malaysia*

*Corresponding author's e-mail: ** ***i.wicaksana001@binus.ac.id***

***Article History:***

*Received: 15th February 2025*
*Revised: 22nd March 2025*
*Accepted: 22nd May 2025*
*Available online: 1st September 2025*

***Keywords:***

*Classification;*
*Convolutional Neural Network;*
*Deep Learning;*
*Skeletal Malocclusion;*
*Vision Attention.*

## *ABSTRACT*

*Skeletal malocclusion, a common orthodontic condition, affects jaw function and dental health. It is often caused by genetic factors, abnormal growth, bad habits, or trauma. Conventional diagnostic models often fail to generalize across diverse datasets, leading to overfitting and poor test performance. This study aimed to improve diagnostic accuracy by incorporating Vision Attention mechanisms into a custom Convolutional Neural Network (CNN), enabling the model to focus on critical regions in X-ray images. A total of 491 radiographic images depicting facial skeletal structures with various malocclusion types (Classes 1, 2, and 3) were used in this study. A custom CNN was developed and evaluated both with and without attention mechanisms—specifically, Scaled Dot Product Attention and Multihead Attention—to assess their impact on classification performance. The baseline CNN without attention achieved an accuracy of 0.68. With Scaled Dot Product Attention, accuracy improved to 0.72, while Multihead Attention achieved the highest accuracy of 0.76. Evaluation using weighted average precision, recall, and F1-score showed that attention mechanisms significantly enhanced the model's ability to differentiate between malocclusion classes. Notably, the Multihead Attention model yielded the best performance, reducing misclassification errors and improving generalization. Confusion matrix analysis revealed that it had the lowest classification errors, especially in distinguishing between Class 0 and Class 1. These results suggest that incorporating attention mechanisms, particularly Multihead Attention, enhances CNN performance by improving feature extraction and classification accuracy. Future research should explore more diverse datasets and implement advanced augmentation techniques to improve clinical reliability.*

***How to cite this article:***

I. P. R. E. Wicaksana, A. Wibowo, Rojali, A. A. Samah, and A. Alias, "CLASSIFICATION OF SKELETAL MALOCCLUSION USING CONVENTIONAL NEURAL NETWORK (CNN) WITH VISION ATTENTION," *BAREKENG: J. Math. & App.*, vol. 19, iss. 4, pp. 2709-2726, December, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article** · **Open Access**

## 1. INTRODUCTION

Malocclusion is a significant dental and jaw health issue, as it can affect oral function and lead to various complications. Globally, its prevalence is estimated at around 56%, with the highest rates reported in Africa (81%) and Europe (72%) [1]. In Indonesia, a study in Yogyakarta found that 57.3% of children suffer from Class I malocclusion, 41.6% from Class II, and 3.3% from Class III, emphasizing the need for appropriate and early treatment [2]. This highlights the importance of accurate and accessible diagnostic tools to support orthodontic care and early intervention.

**Table 1. Malocclusion Characteristics**

| Malocclusion Class | Main Characteristics |
|---|---|
| Class I (Normal/Mild Malocclusion) | Relatively normal bite with slight misalignment of the teeth. The upper teeth slightly overlap the lower teeth, but the relationship between the upper and lower jaws remains normal. |
| Class II (Retrognathic/Overbite) | The upper jaw is positioned ahead of the lower jaw, resulting in an overbite (upper teeth covering most of the lower teeth). It can be divided into Division 1 (protruding upper incisors) and Division 2 (upper incisors tilted inward). |
| Class III (Prognathic/Underbite) | The lower jaw is positioned ahead of the upper jaw, resulting in an underbite (lower teeth in front of the upper teeth). |

Based on Table 1, these malocclusion classes also affect the overall facial profile and jaw positioning. In Class I, there are usually no significant issues with jaw alignment, although dental problems such as crowding or spacing may occur. In Class II cases, the lower jaw appears smaller or recessed, leading to a convex facial profile. Conversely, Class III malocclusion often presents with a prominent lower jaw, resulting in a concave facial appearance or a more pronounced chin. Understanding these distinctions is essential in orthodontic diagnosis and treatment planning [3].

Biologically and medically, these distinctions are critical because each class has different etiologies, implications for facial growth, and treatment strategies. For instance, Class II and III malocclusions may require orthopedic or even surgical intervention, while Class I might be managed with dental braces alone. From a medical imaging perspective, these differences are visible in the relative positioning and angles of the jaw bones and dental arches on lateral skull radiographs. Accurate classification helps clinicians develop appropriate treatment plans, predict growth patterns, and avoid complications in long-term oral health. Therefore, automating this classification using deep learning can support orthodontists in diagnosis and reduce subjectivity in clinical assessments [3].

Deep learning, particularly Convolutional Neural Networks (CNNs), has shown great potential in classifying skeletal malocclusion from radiographic images by automatically extracting features related to jaw positioning and facial anatomy [4], [5]. However, CNNs often struggle to emphasize the most relevant regions in complex medical images, which can result in suboptimal classification accuracy. To overcome this limitation, vision attention mechanisms have been introduced to help the model focus on important image areas while minimizing the influence of irrelevant features [6], [7].

Despite prior successes, existing deep learning models still face challenges such as generalizing across diverse patient data and imaging conditions. Additionally, many models operate as “black boxes”, offering little transparency in how predictions are made. To address these issues, interpretability techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) are employed to visually reveal the image regions that influence model predictions. This not only improves trust in AI-driven diagnostics but also provides clinicians with deeper insights.

This study proposes a skeletal malocclusion classification model based on a custom CNN architecture integrated with Scaled Dot Product Attention and Multihead Attention mechanisms. These attention modules are designed to guide the model in identifying critical regions in radiographic images, thus improving diagnostic precision. By comparing the performance of CNN models with and without attention mechanisms, the research evaluates the effectiveness of attention in enhancing classification accuracy.

Numerous studies have explored CNNs for malocclusion and medical image classification. For example, CNNs were utilized to classify malocclusion from cranio-spinal cephalometric images, both with and without jaw masking [8], and to predict orthognathic surgery needs in patients with dentofacial dysmorphisms using cephalograms [9]. DenseNet161, ResNet152, GoogleNet, and VGG16 have been applied to automatically classify sagittal craniofacial patterns, with DenseNet161 reportedly performing best, although clear accuracy metrics were not provided [10]. Additionally, CNNs with Grad-CAM were used to predict mandibular growth trends in children with anterior crossbite [11], and the cut-out method was implemented to focus on incisors in malocclusion classification, though results were not optimal [3]. CNNs have also been widely used in broader medical imaging tasks, such as brain tumor classification using MRI images, demonstrating their versatility in handling complex visual data [12].

Hybrid methods combining CNNs with other machine learning models have also been proposed. An ensemble approach was used to detect periodontal disease and malocclusion, although details on data evaluation were limited [13]. CNNDWNet, combined with Gradient Boosting, was applied for COVID-19 and pneumonia classification from CT and X-ray images [14]. These studies highlight the potential of hybrid models, but many lacked transparency in evaluation or failed to demonstrate superior performance over standalone CNNs.

Outside of CNNs, other studies used rule-based and machine learning algorithms for early prediction of Class III malocclusion from profile images, with traditional machine learning performing better than deep learning in that specific context [15]. Heuristic methods were also employed to detect Class III malocclusion using facial landmarks, but their performance varied across skeletal classes [16]. Another work classified sagittal spinal patterns in children using lateral cephalograms and profile photos [17].

Although previous works have demonstrated the utility of CNNs in malocclusion detection, several challenges remain. Many models struggle with generalization, lack interpretability, and do not leverage attention mechanisms to focus on critical image regions. Moreover, few studies evaluate or compare the impact of integrating attention into CNN architectures. Our research addresses these gaps by proposing a Custom CNN enhanced with Scaled Dot Product Attention and Multihead Attention. We explicitly assess their effects on classification performance and interpretability using tools like Grad-CAM, providing a more transparent and clinically relevant diagnostic model.

## 2. RESEARCH METHODS

### 2.1 Research Stages

As shown in Figure 1, the first stage involves identifying relevant data sources and collecting or downloading the data, as well as checking the authenticity and quality of the data. Next, the data is processed through pre-processing to clean, tidy, and prepare the data for further analysis. The next stage is model building, which involves building a model or algorithm that is appropriate to the research being conducted. In the context of CNN and vision attention, this involves creating a CNN architecture that fits skeletal malocclusion data, as well as using vision attention techniques to improve model performance. After the model is created, the next step is training the model on previously processed data. The model validation stage is then carried out by testing the model on data that has never been seen before to measure model performance objectively. This also involves a model optimization process, if necessary. Model evaluation is carried out to analyze the results of the model that has been trained and validated, including identifying the strengths, weaknesses, and limitations of the model developed.

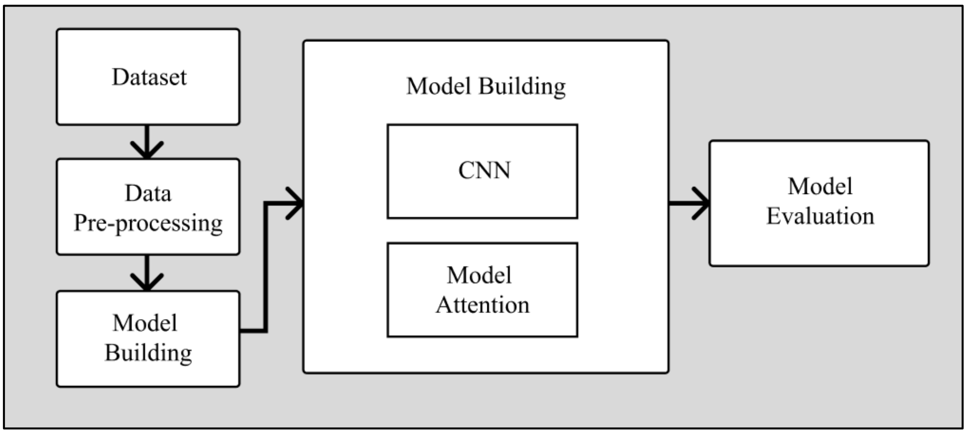


**Figure 1. Research Stages**

### 2.2 Dataset

The dataset, as shown in Figure 2, used in this study was originally collected by researchers from Universiti Teknologi Malaysia (UTM). It comprises 491 radiographic images of facial skeletal structures categorized into three malocclusion classes: Class I with 240 images, Class II with 184 images, and Class III with 67 images. This study utilizes the dataset solely for training and evaluating the proposed classification model.

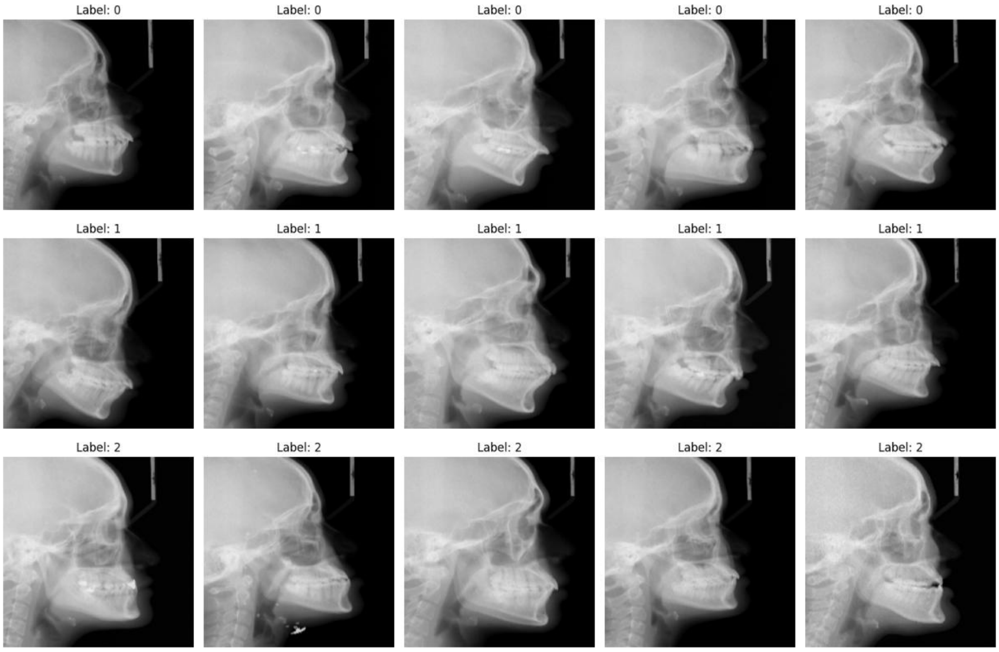


Figure 2. **Dataset Malocclusion**

### 2.3 Data Pre-Processing

Data pre-processing ensures that the dataset is of high quality and ready for CNN training and testing. Several techniques are used in data processing. First, image loading and normalization are performed to transform the images into a suitable format for processing by the Convolutional Neural Network (CNN). Normalization is applied to scale pixel values between 0 and 1, which enhances model stability and accelerates convergence during training. Subsequently, cropping and resizing techniques are employed to remove irrelevant parts of the images while maintaining the critical regions that contribute to classification accuracy. After cropping, the images are resized to a standard dimension to ensure uniformity across the dataset.

Categorical labels are converted into one-hot encoded format. This transformation ensures that the model can effectively process multi-class classification tasks by representing each class as a separate binary vector. Next, the dataset is split into training, validation, and testing subsets with an 80:10:10 ratio. This ensures that the model is trained on a portion of the data while another portion remains unseen for evaluation, allowing an assessment of the model's ability to generalize beyond the training data. Oversampling techniques are applied to address class imbalance issues. This process involves duplicating samples from underrepresented classes to achieve a balanced distribution, thereby preventing the model from becoming biased toward majority classes.

All training and experiments were performed using the free tier of Google Colaboratory, a cloud-based development environment equipped with GPU acceleration. The implementation was done in Python using TensorFlow and Keras, along with libraries such as NumPy, Pandas, scikit-learn, Matplotlib, and Seaborn.

### 2.4 CNN Custom

The main structure of a CNN consists of several layers, including convolutional layers responsible for performing convolution operations on the input image, followed by other layers such as pooling layers and fully connected layers [18], [19]. The application process involves loading pre-processed images (resized, cleaned, and augmented) into the model, which represents them as 3D tensors. These tensors contain height, width, and color dimensions.

The convolutional layers apply a series of filters to extract basic visual features such as edges, textures, and colors. After convolution, the ReLU activation function introduces non-linearity by setting negative values to zero, helping the model capture complex relationships between features and preventing vanishing gradients. Max pooling reduces image dimensions by selecting the maximum value from sub-areas, retaining essential information.

Once features are extracted, they are flattened into a 1D tensor for the fully connected layer, which learns the non-linear relationships between features and class labels. During training, the model minimizes prediction errors by optimizing parameters using the Adam optimizer, which adjusts learning rates for each parameter to reach convergence quickly and stably. After training, the model is used to make predictions on new images, and performance is evaluated using metrics such as accuracy, precision, recall, and F1-score [20].

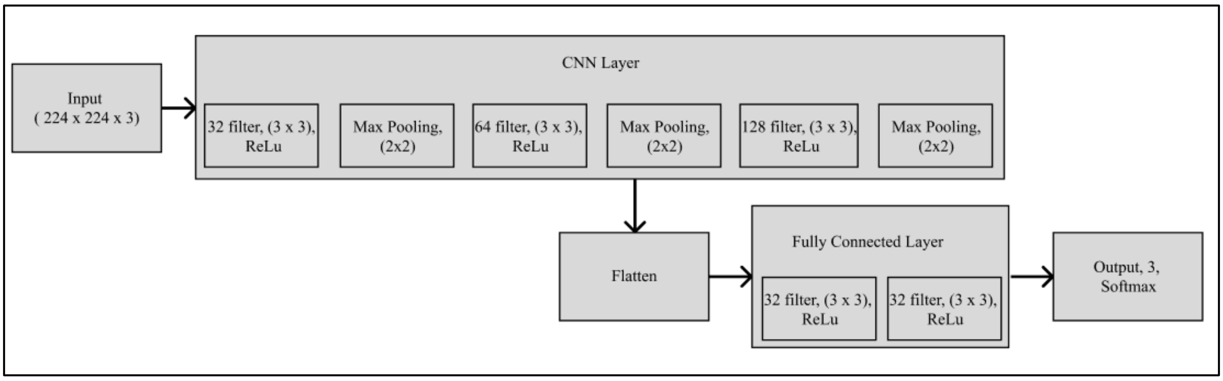


**Figure 3. Architecture of CNN**

As shown in Figure 3 and Table 2, this CNN architecture is designed to accept color images with dimensions of 224×224×3, meaning that each image has a height of 224 pixels, a width of 224 pixels, and 3 color channels (RGB). This input serves as a starting point in processing the features that will be extracted by the subsequent layers. This architecture has three main blocks consisting of convolutional layers.

**Table 2. Summary Table of Architecture CNN**

| Layer (Type) | Output Shape | Param |
|---|---|---|
| InputLayer | (None, 224, 224, 3) | 0 |
| Conv2D (32 filters) | (None, 224, 224, 32) | 896 |
| MaxPooling2D | (None, 112, 112, 32) | 0 |
| Conv2D (64 filters) | (None, 112, 112, 64) | 18,496 |
| MaxPooling2D | (None, 56, 56, 64) | 0 |
| Conv2D (128 filters) | (None, 56, 56, 128) | 73,856 |
| MaxPooling2D | (None, 28, 28, 128) | 0 |
| Flatten | (None, 100352) | 0 |
| Dense (128 units) | (None, 128) | 12,845,184 |
| Dropout | (None, 128) | 0 |
| Dense (Output, 3 classes) | (None, 3) | 387 |

The first block consists of a single convolution layer with 32 3×3 filters, using the ReLU activation function and “same” padding. The main purpose of this convolution layer is to extract basic features from the input image, such as edges and textures. After that, a max pooling operation with a size of 2×2 is performed to reduce the spatial dimension and retain important information. The second block follows the same pattern but with an increased number of filters to 64. The increase in the number of filters aims to capture more complex features, such as broader shape and contour patterns. The third block has a convolution layer with 128 filters, which allows for the refinement of more abstract high-level features from the image. After this convolution process, max pooling is applied again to reduce the dimensionality and increase computational efficiency. The convolutional layers in this architecture play a role in extracting features with increasing complexity as the network deepens. Meanwhile, the pooling layers help reduce the number of parameters and prevent overfitting by retaining the most significant information from the extracted features.

After going through a series of convolution and pooling layers, the extracted features are flattened into a one-dimensional vector through a flattening process. This vector is then passed to a fully connected layer with 128 neurons and a ReLU activation function. This layer is responsible for connecting the previously extracted features into a more meaningful representation for the classification task. To reduce the risk of overfitting, a 50% dropout is applied, which randomly deactivates half of the neurons during the training process. This technique improves the generalization of the model by reducing over-reliance on a particular subset of neurons. The last layer consists of several neurons corresponding to the number of classes in the classification task. Each neuron in this layer uses a softmax activation function, which transforms the output into a probability distribution. The softmax function ensures that the sum of the probabilities of all classes is 1, allowing the model to produce predictions that can be interpreted well.

## 2.5 Scaled Dot Product Attention

Vision Scaled Dot Product Attention is a mechanism used in image-processing models like Transformers for tasks such as image segmentation and restoration. This attention mechanism helps the model focus on important parts of the image during processing. As shown in **Figure 4** and **Table 3**, attention is applied before the flattened data is processed in the fully connected layer.

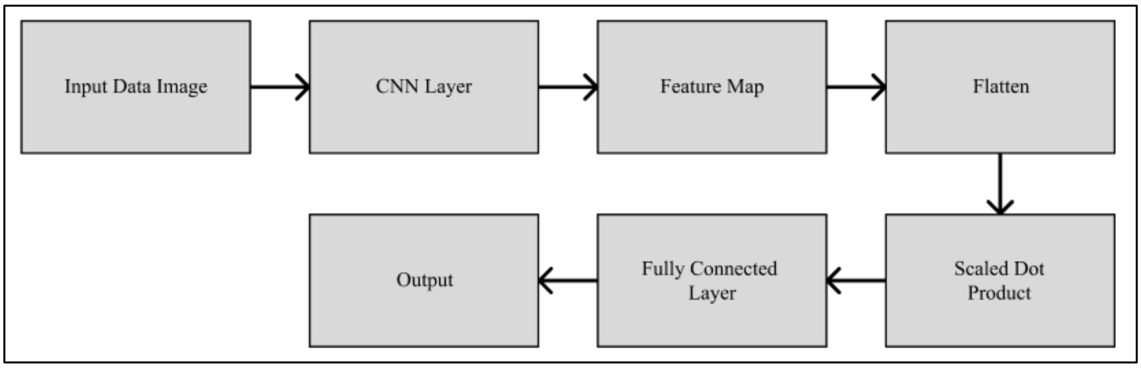


**Figure 4. CNN Architecture with Scaled Dot Product Attention**

**Table 3** shows the architecture summary of the CNN model enhanced with Scaled Dot Product Attention. The model begins with three convolutional blocks (Conv2D + MaxPooling), gradually increasing the number of filters (32, 64, 128), followed by flattening the feature maps. Three parallel Dense layers generate the Query, Key, and Value vectors, which are fed into the Scaled Dot Product Attention mechanism. The attention output is then concatenated with the original flattened features to enrich contextual understanding. This combined feature vector is passed through a Dense layer, dropout for regularization, and finally a softmax output layer for 3-class classification. The architecture integrates attention to enhance feature weighting and interpretability while maintaining the CNN backbone for spatial feature extraction.

**Table 3. Summary Table of CNN Architecture with Scaled Dot Product**

| Layer (Type) | Output Shape | Param | Connected to |
|---|---|---|---|
| InputLayer | (None, 224, 224, 3) | 0 | - |
| Conv2D (32 filters) | (None, 224, 224, 32) | 896 | InputLayer |
| MaxPooling2D | (None, 112, 112, 32) | 0 | Conv2D |
| Conv2D (64 filters) | (None, 112, 112, 64) | 18,496 | MaxPooling2D |
| MaxPooling2D | (None, 56, 56, 64) | 0 | Conv2D |
| Conv2D (128 filters) | (None, 56, 56, 128) | 73,856 | MaxPooling2D |
| MaxPooling2D | (None, 28, 28, 128) | 0 | Conv2D |
| Flatten | (None, 100352) | 0 | MaxPooling2D |
| Dense (128 units) x3 | (None, 128) | 38,535,552 | Flatten (shared connection) |
| ScaledDotProductAttention | (None, 128) | 0 | Dense x3 |
| Concatenate | (None, 100480) | 0 | Flatten + Attention Output |
| Dense (128 units) | (None, 128) | 12,861,568 | Concatenate |
| Dropout | (None, 128) | 0 | Dense |
| Dense (Output, 3 classes) | (None, 3) | 387 | Dropout |

As shown in Figure 5 and Equation (1), the attention process involves receiving input features from the CNN, which are then projected into three different matrices: Query (Q), Key (K), and Value (V), where $\sqrt{d_k}$ is the dimension of the query and key vectors, $softmax$ is the function calculating the weights, $T$ is the transpose operation, and $\sqrt{d_k}$ is a scaling factor to stabilize calculations [7].

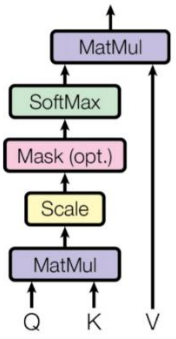


**Figure 5. Scaled Dot Product Attention Mechanism [7]**

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where:

$Q$ : Query matrix
$K$ : Key matrix
$V$ : Value matrix
$d_k$ : Dimension of the key vectors
$QK^T$ : Dot product between query and transposed key
$softmax$ : Normalizes scores to probabilities

The query focuses on specific parts of the image, the key identifies correlations between features, and the value stores the relevant information. The dot product of the query and key is scaled by the square root of the key's dimension to stabilize the output, followed by the application of softmax to compute attention weights. These weights are then multiplied by the value to produce the attention output, highlighting the important parts of the image [7].

## 2.6 Multihead Attention

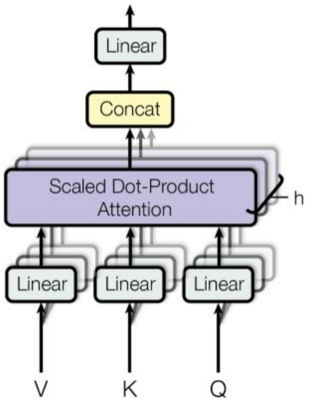


**Figure 6. Multihead Attention Mechanism [7]**

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h)W^O \tag{2}$$
$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

Where:

$head_i$ : Output of the ith attention head

$h$ : Total number of heads
$W^O$ : Output projection matrix
$Concat$ : Concatenation operation over all heads
$W_i^Q, W_i^K, W_i^V$ : Linear projection matrices of queries, keys, and values for $head_i$
$Q, K, V$ : Query, Key, Value matrices

In the Vision Multihead Attention mechanism, shown in Figure 6 and Equation (2) and Equation (3), multiple sets of queries, keys, and values are created, with each set processed in parallel through different linear projections. Each set, or "head," independently runs scaled dot-product attention, where the dot product between query and key is computed, scaled, and passed through softmax to calculate attention weights. After processing, the output from each head is concatenated into a single representation, integrating various perspectives obtained from each head [7].

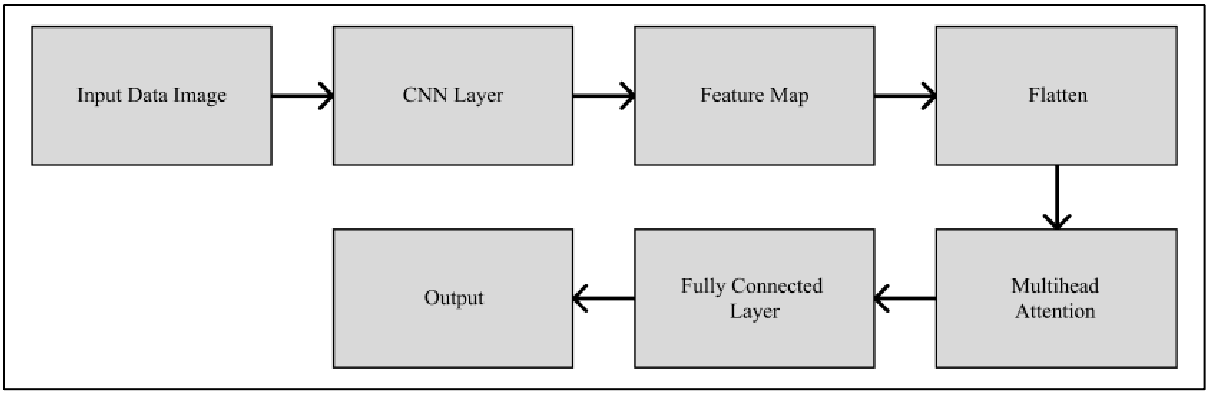


**Figure 7. CNN Architecture with Multihead Attention**

This final representation is then projected back through linear projection to produce the final output. By simultaneously focusing on different subspaces of the input data, Multihead Attention enhances the model’s ability to capture complex data patterns. Shown in Figure 7 and Table 4 after processing by the multihead layer, features are combined and processed by fully connected and output layers, allowing the CNN model to classify skeletal malocclusion based on the extracted and processed features.

**Table 4. Summary Table of the Multihead Attention Architecture**

| Layer (Type) | Output Shape | Param | Connected to |
|---|---|---|---|
| InputLayer | (None, 224, 224, 3) | 0 | - |
| Conv2D (32 filters) | (None, 224, 224, 32) | 896 | InputLayer |
| MaxPooling2D | (None, 112, 112, 32) | 0 | Conv2D |
| Conv2D (64 filters) | (None, 112, 112, 64) | 18,496 | MaxPooling2D |
| MaxPooling2D | (None, 56, 56, 64) | 0 | Conv2D |
| Conv2D (128 filters) | (None, 56, 56, 128) | 73,856 | MaxPooling2D |
| MaxPooling2D | (None, 28, 28, 128) | 0 | Conv2D |
| Flatten | (None, 100352) | 0 | MaxPooling2D |
| Dense (128 units) | (None, 128) | 12,845,184 | Flatten |
| Reshape x3 | (None, 1, 128) | 0 | Dense |
| MultiHeadAttention | (None, 1, 128) | 263,808 | Reshape x3 |
| Flatten | (None, 128) | 0 | MultiHeadAttention |
| Concatenate | (None, 100480) | 0 | Flatten (CNN) + Flatten (MHA) |
| Dense (128 units) | (None, 128) | 12,861,568 | Concatenate |
| Dropout | (None, 128) | 0 | Dense |
| Dense (Output, 3 classes) | (None, 3) | 387 | Dropout |

### 2.7 Training Configuration

Table 5. **Training Configuration**

| Item | Value / Setting |
|---|---|
| Optimizer | Adam (learning rate = 0.0001) |
| Loss Function | Categorical Crossentropy |
| Metrics | Accuracy |
| Epochs | 70 |
| Batch Size | 64 |
| Early Stopping | patience=5 |
| Input image size | 224 x 224 x 3 |

As shown in **Table 5**, the models are trained using the Adam optimizer with a learning rate of 0.0001. Adam is an efficient optimizer that helps the model learn quickly and smoothly. A small learning rate ensures the model doesn't overshoot and learns in small, stable steps. The loss function used is categorical cross-entropy, which is suitable for classification tasks with more than two categories. The accuracy metric is used to measure how often the model predicts the correct class.

The training process runs for up to 70 full passes (epochs) over the training data. However, an early stopping mechanism is included to stop the training early if the model's performance on the validation data does not improve for 5 consecutive epochs, helping prevent overfitting. The training data is divided into smaller batches of 64 images at a time. This is called the batch size and helps make training more efficient and stable. The validation data is prepared as a separate dataset (not split automatically), so the model can be evaluated on unseen data during training. Each input image is resized to 224 x 224 pixels with 3 color channels (RGB) — a standard size that balances detail and training speed, commonly used in CNN models like ResNet.

### 2.8 Model Comparison

After the evaluation of CNN models in classifying skeletal malocclusion using vision attention-based CNNs with image data is conducted using several common evaluation metrics for classification, such as accuracy, recall, precision, F1-score, and the confusion matrix.

The confusion consists of:

| | |
|---|---|
| True Positive (TP) | : correctly predicted as a positive class. |
| False Positive (FP) | : incorrectly predicted as a positive class. |
| False Negative (FN) | : incorrectly predicted as a negative class. |
| True Negative (TN) | : correctly predicted as a negative class. |

This evaluation helps examine the performance of the CNN and vision attention models in classifying radiographic images of skeletal malocclusion into the correct classes. Accuracy measures the ratio of correct predictions to the total number of data points. Precision measures how accurate the model is in identifying the positive class. Recall measures how well the model detects the positive class. F1-score is the mean of precision and recall, balancing both metrics. In this study, we use the weighted average (weight average) for precision, recall, and F1-score to account for class imbalances. The weighted average ensures that the contribution of each class to the final score is proportional to the number of samples in that class, preventing classes with fewer samples from disproportionately affecting the evaluation. This approach provides a more reliable overall assessment of the model's performance across all classes.

Accuracy is the ratio of the number of correct predictions to the total amount of data. This metric gives an idea of how well the CNN model is at classifying skeletal malocclusion radiographic images. This accuracy can be found by using the formula:

$$Accuracy = \frac{\sum_{i=1}^{k} TPi}{\sum_{i=1}^{k} (TPi + FPi + FNi)} \tag{4}$$

where:

$TP_i$ : True Positives for class $i$
$FP_i$ : False Positives for class $i$
$FN_i$ : False Negative for class $i$
$k$ : Number of classes

Precision is the ratio between the number of positive classes that are correctly predicted and the total number of positive classes predicted by the model. This metric gives an idea of how accurate the model is in identifying skeletal malocclusion classes. Precision can be found using the formula:

$$Precision\ (i) = \frac{TPi}{TPi + FPi} \tag{5}$$

where:

$TP_i$ : True Positives for class $i$
$FP_i$ : False Positives for class $i$

Recall is the ratio of the number of positive classes that are correctly predicted to the total number of positive classes that should be present. This metric gives an idea of how well the model can recognize skeletal malocclusion classes. This recall can be found by using the formula:

$$Recall\ (i) = \frac{TPi}{TPi + FNi} \tag{6}$$

where:

$TP_i$ : True Positives for class $i$
$FN_i$ : False Negatives for class $i$

F1-score is the harmonic mean of precision and recall. This metric gives an idea of how well the model can maintain a balance between precision and recall. The F1-score can be found using the formula:

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

where:

$Recall(i)$ : Recall for class $i$
$Precision(i)$ : Precision for class $i$

## 3. RESULTS AND DISCUSSION

As seen in the training and validation curves for the CNN model, as shown in Figure 8 and Figure 9, the performance dynamics throughout the epochs provide insight into their learning behavior and generalization ability. The baseline CNN model exhibited a steady increase in training accuracy, reaching approximately 80% by the final epoch. However, its validation accuracy plateaued around 60–65% and fluctuated after the 15th epoch. This gap between training and validation performance indicates overfitting, where the model memorizes training data but struggles to generalize to unseen data. The training loss decreased significantly throughout the epochs, but the validation loss stopped improving and slightly increased, further confirming the presence of overfitting in the baseline model.

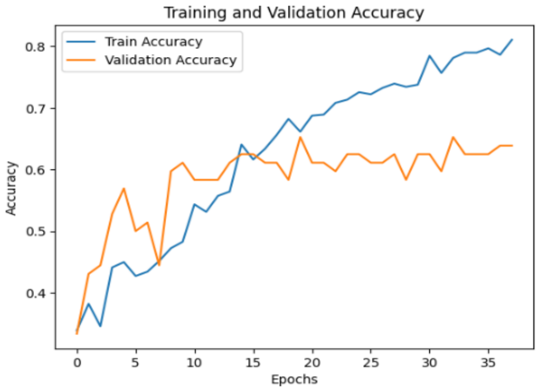


**Figure 8. Training and Validation Accuracy of the CNN Model**

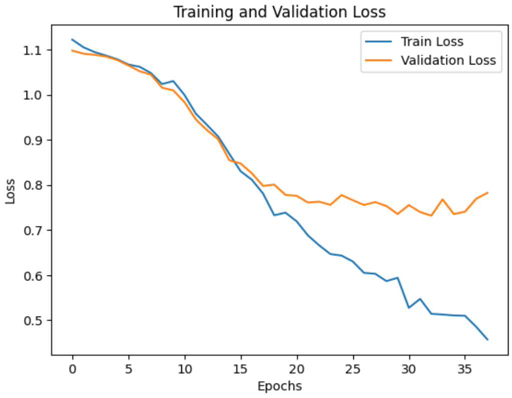


**Figure 9. Training and Validation Loss of the CNN Model**

In comparison shown in **Figure 10** and **Figure 11**, the CNN model enhanced with Scaled Dot Product Attention showed improved learning stability. Training accuracy steadily rose to about 78%, while validation accuracy consistently increased to approximately 70% without drastic fluctuations. Both training and validation loss declined in parallel throughout the training, suggesting that this attention mechanism helped reduce overfitting and improved generalization. The model was better able to capture relevant feature relationships, leading to a more balanced and reliable performance.

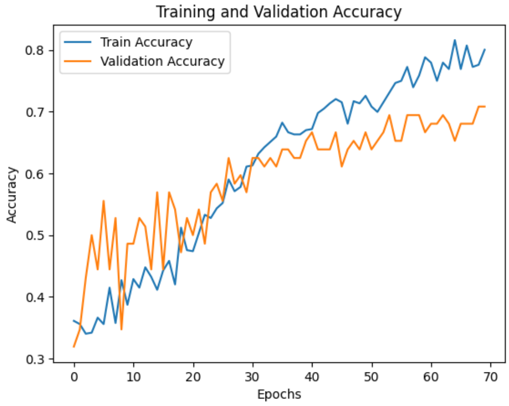


**Figure 10. Training and Validation Accuracy of the CNN Model with Scaled Dot Product**

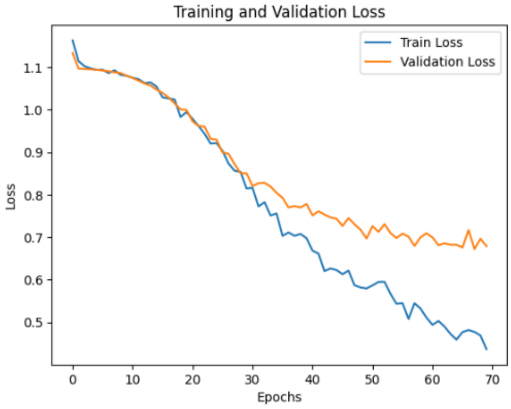


**Figure 11. Training and Validation Loss of the CNN Model with Scaled Dot Product**

The best results were observed in the CNN model incorporating Multihead Attention, shown in **Figure 12** and **Figure 13**, which achieved the highest training accuracy of around 92% and a validation accuracy of nearly 78%. The smaller gap between training and validation accuracy suggests strong generalization ability. The loss curves also support this, as both training and validation loss decreased significantly, with only minor fluctuations near the end. These results highlight that the Multihead Attention mechanism enables the model to learn richer feature representations, ultimately leading to superior classification performance.

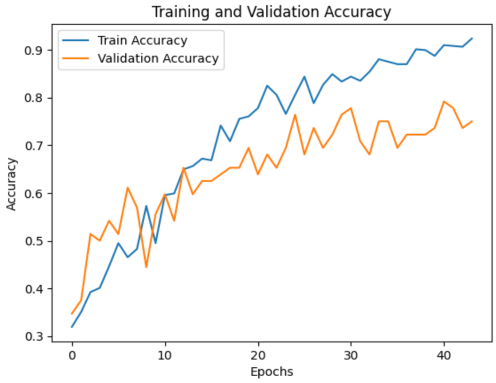


**Figure 12. Training and Validation Accuracy of the CNN Model with Multihead Attention**

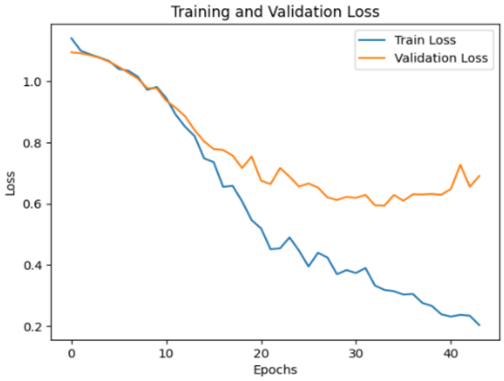


**Figure 13. Training and Validation Loss of the CNN Model with Multihead Attention**

The performance of the three models—CNN, CNN with Scaled Dot Product, and CNN with Multihead Attention—was evaluated based on the confusion matrix and classification report, as shown in **Figure 14**, **Figure 15**, **Figure 16**, and **Table 6**. The primary objective of this study was to enhance the classification accuracy of skeletal malocclusion.

The baseline CNN model achieved an accuracy of 68%, the lowest among the three models. The addition of the Scaled Dot Product Attention mechanism improved the accuracy to 72%, indicating that this technique helps the model better capture feature relationships. The best-performing model was the CNN with Multihead Attention, achieving an accuracy of 76%, demonstrating that this mechanism provides richer feature representations and improves generalization.

In precision analysis, the CNN model with Multihead Attention showed the highest precision, achieving an average weighted precision of 0.76. This indicates that it is more effective in reducing false positives compared to the other models. Similarly, for recall, the CNN with Multihead Attention outperformed the other models, suggesting its superior capability in recognizing samples across all classes.

The F1-score, which balances precision and recall, showed that the CNN model with Multihead Attention consistently performed the best across all classes. The improved balance between precision and recall highlights the model's ability to detect samples with fewer errors.

From the confusion matrices, the baseline CNN model exhibited a higher number of misclassifications, particularly between classes 0 and 1, showing that it struggles to distinguish between these two categories. The CNN model with Scaled Dot Product Attention reduced misclassification in class 2 but still had errors in class 1. The CNN model with Multihead Attention demonstrated the best prediction distribution, minimizing classification errors across all classes. This suggests that the Multihead Attention mechanism plays a crucial role in improving classification performance and reducing errors.

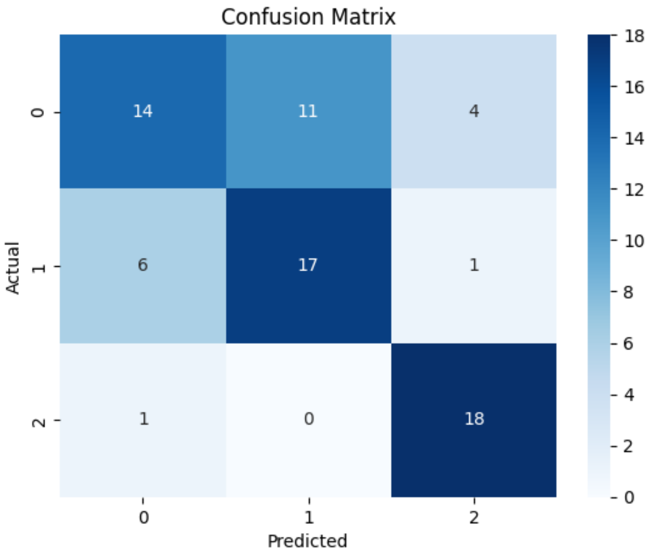


**Figure 14. Confusion Matrix of the CNN Model**

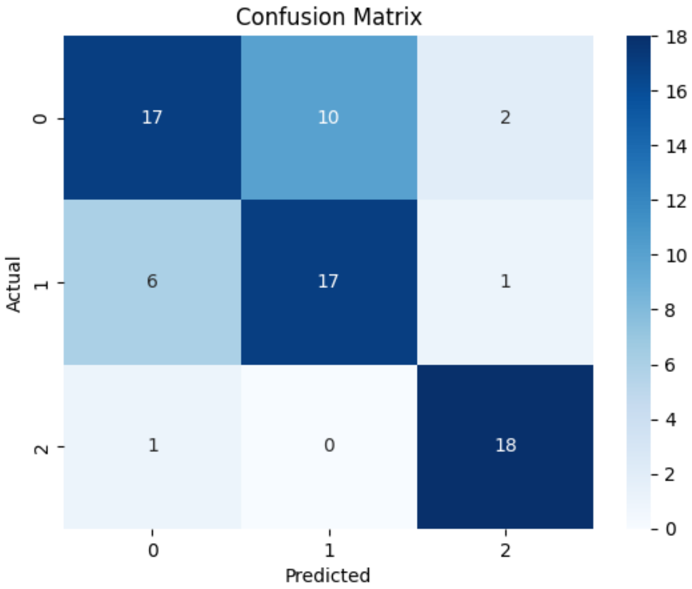


**Figure 15. Confusion Matrix of the CNN Model with Scaled Dot Product Attention**

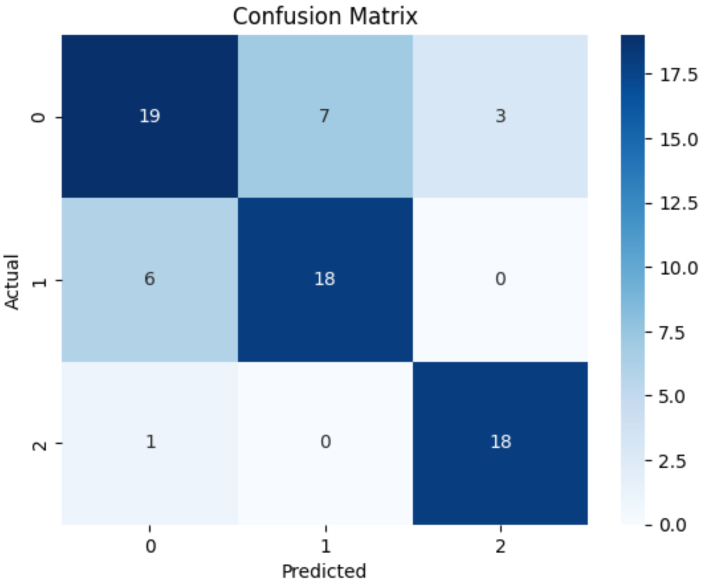
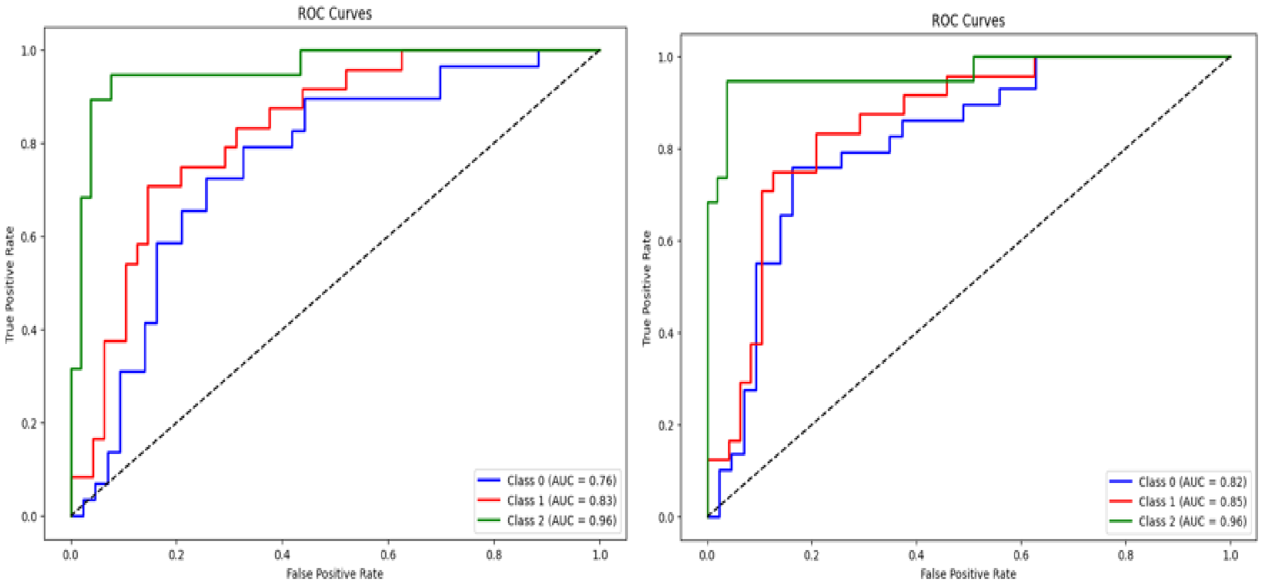


**Figure 16. Confusion Matrix of the CNN Model with Multihead Attention**

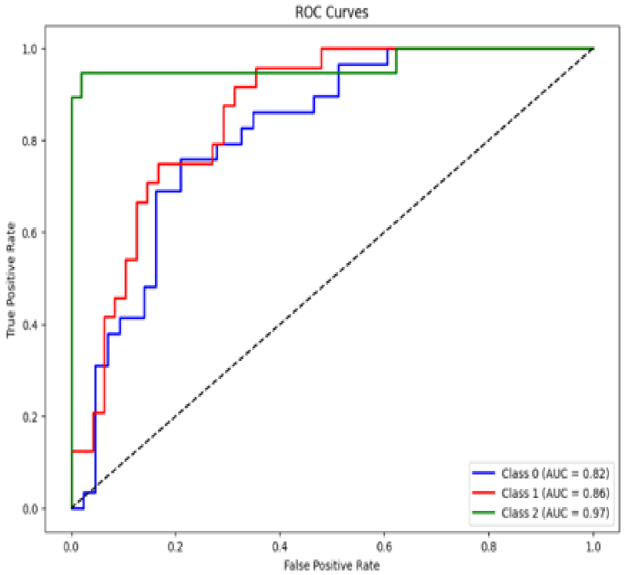
**Table 6. Classification Report**

| Metric | CNN | CNN+Scaled Dot Product | CNN+Multihead Attention |
|---|---|---|---|
| Accuracy | 0.68 | 0.72 | 0.76 |
| Avg Weight Precision | 0.68 | 0.72 | 0.76 |
| Avg Weight Recall | 0.68 | 0.72 | 0.76 |
| Avg Weight F1-Score | 0.67 | 0.72 | 0.76 |

**Table 6** shows that the CNN + Multihead Attention model achieves the highest accuracy (0.76), outperforming both the baseline CNN model (0.68) and the CNN + Scaled Dot Product Attention model (0.72). This improvement can be attributed to the ability of Multihead Attention to capture multiple representation subspaces simultaneously, allowing the model to focus on different parts of the image in parallel. Unlike standard Scaled Dot Product Attention, which uses a single attention mechanism, Multihead Attention provides richer contextual understanding by combining diverse attention outputs. This leads to a more robust feature representation, especially in complex classification tasks like skeletal malocclusion, where fine-grained spatial patterns across facial structures are crucial for accurate prediction. Consequently, the model becomes better at distinguishing subtle variations between malocclusion classes, resulting in improved overall performance.



**(a)** **(b)**



**(c)**

**Figure 17. ROC-AUC Result from (a) CNN, (b) CNN with Scaled Dot Product, and (c) CNN with Multihead Attention**

As shown in **Figure 17** and **Table 7**, the ROC-AUC results for malocclusion classification using three different models: standard CNN, CNN with Scaled Dot Product Attention, and CNN with Multihead Attention, demonstrate strong overall performance, particularly for Class 2, which consistently achieved high AUC scores above 0.96 across all models. The CNN struggled with Class 0, showing the lowest AUC of 0.76, indicating difficulty in distinguishing that class. The incorporation of attention mechanisms improved the model's ability to handle more subtle features, as seen in the Scaled Dot Product Attention model, which raised Class 0 and Class 1 performance. The CNN with Multihead Attention achieved the best overall results, with AUC scores of 0.82, 0.86, and 0.97 for Class 0, Class 1, and Class 2, respectively. These findings suggest that attention mechanisms, particularly multihead attention, enhance the model's discriminative capacity across all classes, especially for those previously harder to classify.

**Table 7. ROC-AUC Result**

| Model | Class 0 | Class 1 | Class 2 |
|---|---|---|---|
| CNN | 0.76 | 0.83 | 0.96 |
| CNN + Scaled Dot Product | 0.82 | 0.85 | 0.96 |
| CNN + Multihead Attention | 0.82 | 0.86 | 0.97 |

The Grad-CAM visualizations across the three models—namely the baseline CNN, CNN with Scaled Dot Product Attention, and CNN with Multihead Attention—demonstrate a progressive enhancement in model interpretability. The baseline CNN exhibits relatively dispersed activation regions, indicating less precision in localizing diagnostically relevant features within the cephalometric X-ray images. By incorporating Scaled Dot Product Attention, the model shows improved focus on key anatomical structures, reflecting an increased ability to capture salient patterns associated with skeletal malocclusion. The CNN model equipped with Multihead Attention further refines this capability, generating more concentrated and consistent heatmaps. This suggests that the model is able to attend to multiple informative regions simultaneously, enhancing both its predictive performance and explainability. These findings indicate that the integration of attention mechanisms can significantly improve the transparency and diagnostic relevance of deep learning models in medical image classification tasks.

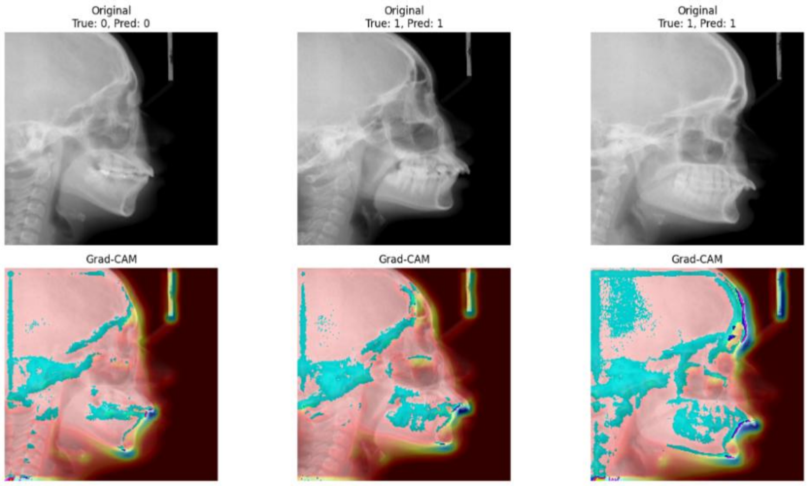


**Figure 18. Visualization Grad-Cam CNN**

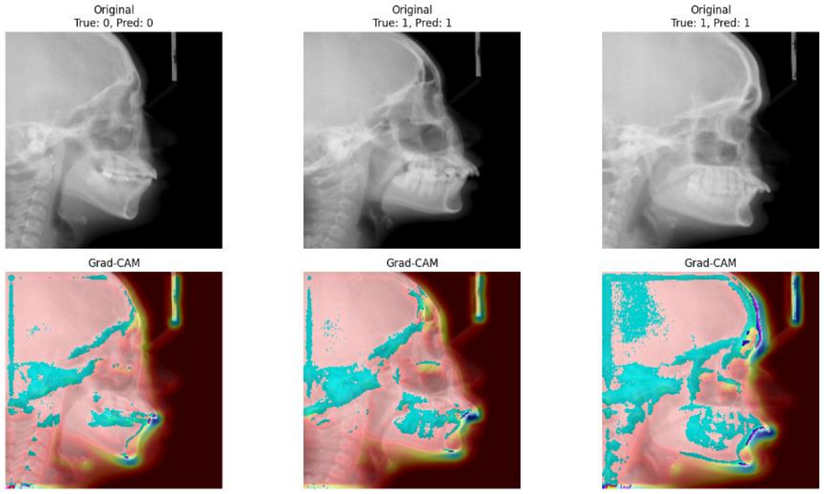


**Figure 19. Visualization Grad-Cam CNN with Scaled Dot Product**

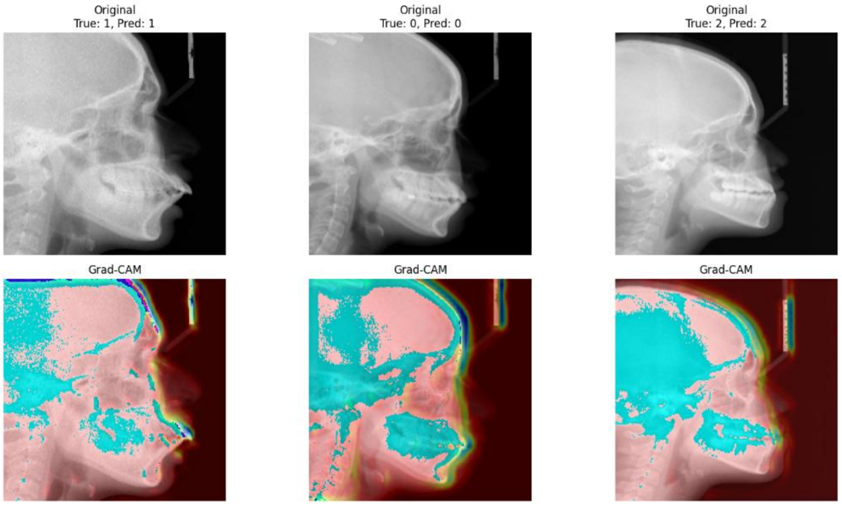


**Figure 20. Visualization Grad-Cam CNN with Multihead Attention**

# 4. CONCLUSION

## 4.1 Result Summary

From the experimental results, it can be concluded that incorporating attention mechanisms into the custom CNN architecture significantly improves the model's performance in classifying skeletal malocclusion images. The CNN model integrated with Multihead Attention (4 heads) achieved the highest accuracy of 0.76, outperforming the CNN with Scaled Dot Product Attention (0.72) and the baseline CNN without attention (0.68). These results indicate that attention mechanisms enhance the model's ability to focus on important features, leading to better classification performance. Additionally, the use of attention was shown to increase the generalization ability of the model, as reflected in improved validation accuracy.

## 4.2 Limitations

Despite these improvements, the overall accuracy remains relatively modest at 0.76. This suggests that the model still struggles with certain classification tasks, particularly in classes with fewer training samples. A key limitation is the limited size and imbalance of the dataset despite using an oversampling method to fix the inbalance data, which negatively impacts the model's ability to learn representative features across all classes. Furthermore, the architecture used in this study was relatively simple, and there was limited exploration of advanced feature extraction or optimization strategies, such as using augmented data.

## 4.3 Future Research Directions

Future studies should aim to address the data imbalance and size constraints by expanding the dataset and ensuring a more uniform class distribution. Exploring more advanced CNN architectures or integrating attention mechanisms with pre-trained models (such as Vision Transformers or hybrid CNN-Transformer models) could further enhance classification accuracy. Additionally, employing more sophisticated data augmentation techniques and tuning hyperparameters—such as learning rate, number of filters, and kernel size—may improve model robustness. Exploring transfer learning or adversarial training could also lead to performance gains.

# AUTHOR CONTRIBUTIONS

Putu Ronny Eka Wicaksana: Visualization, Methodology, Formal Analysis, Software, Writing Original Draft. Antoni Wibowo: Conceptualization, Methodology, Formal Analysis, Supervision, Writing Review and Editing. Rojali: Conceptualization, Methodology, Formal Analysis, Supervision. Azurah A Samah: Resource, Data curation. Aspalilah Alias: Resource, Data curation. All authors discussed the results and contributed to the final manuscript.

# FUNDING STATEMENT

This research was financially supported by the Project International BINUS Collaboration in partnership with Universiti Teknologi Malaysia (UTM). For further information regarding the funding, please contact the project leader, Dr. Antoni Wibowo, at +62 858-0647-0667 or via email at anwibowo@binus.edu.

# ACKNOWLEDGMENT

This study is a collaborative effort among Bina Nusantara University, Universiti Sains Islam Malaysia, and Universiti Teknologi Malaysia. We would like to sincerely thank everyone involved in this research. Our gratitude extends to the Bina Nusantara University Grant Program for their support and funding, which has been instrumental in facilitating this work. The contributions and cooperation from all parties have been essential to the successful completion of this study.

## CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest associated with this research.

## REFERENCES

[1] S. I. Stomatologic, "WORLDWIDE PREVALENCE OF MALOCCLUSION IN THE DIFFERENT STAGES OF DENTITION: A SYSTEMATIC REVIEW AND META-ANALYSIS," *Eur J Paediatr Dent*, vol. 21, p. 115, 2020. doi: 10.23804/ejpd.2020.21.02.05.

[2] W. Farani and A. MI, "PREVALENSI MALOKLUSI ANAK USIA 9-11 TAHUN DI SD IT INSAN UTAMA YOGYAKARTA," *Insisiva Dental Journal: Majalah Kedokteran Gigi Insisiva*, vol. 10, no. 1, pp. 26–31, 2021. doi: https://doi.org/10.18196/di.v10i1.7534.

[3] M. F. Harun *et al.*, "INCISOR MALOCCLUSION USING CUT-OUT METHOD AND CONVOLUTIONAL NEURAL NETWORK," *Progress in Microbes and Molecular Biology*, 2022. doi: https://doi.org/10.36877/pmmb.a0000279.

[4] M.-H. Guo *et al.*, "ATTENTION MECHANISMS IN COMPUTER VISION: A SURVEY," *Comput Vis Media (Beijing)*, vol. 8, no. 3, pp. 331–368, 2022. doi: https://doi.org/10.1007/s41095-022-0271-y.

[5] L. Cai, J. Gao, and D. Zhao, "A REVIEW OF THE APPLICATION OF DEEP LEARNING IN MEDICAL IMAGE CLASSIFICATION AND SEGMENTATION," *Ann Transl Med*, vol. 8, no. 11, 2020. doi: https://doi.org/10.21037/atm.2020.02.44.

[6] I. R. Ward, H. Laga, and M. Bennamoun, "RGB-D IMAGE-BASED OBJECT DETECTION: FROM TRADITIONAL METHODS TO DEEP LEARNING TECHNIQUES RGB-D IMAGE ANALYSIS AND PROCESSING," 2019. doi: https://doi.org/10.1007/978-3-030-28603-3_8

[7] A. Vaswani *et al.*, "ATTENTION IS ALL YOU NEED," 2023. doi: 10.48550/arXiv.1706.03762.

[8] S. H. Jeong, J. P. Yun, H.-G. Yeom, H. K. Kim, and B. C. Kim, "DEEP-LEARNING-BASED DETECTION OF CRANIO-SPINAL DIFFERENCES BETWEEN SKELETAL CLASSIFICATION USING CEPHALOMETRIC RADIOGRAPHY," *Diagnostics*, vol. 11, no. 4, p. 591, 2021. doi: https://doi.org/10.3390/diagnostics11040591.

[9] W. Shin *et al.*, "DEEP LEARNING BASED PREDICTION OF NECESSITY FOR ORTHOGNATHIC SURGERY OF SKELETAL MALOCCLUSION USING CEPHALOGRAM IN KOREAN INDIVIDUALS," *BMC Oral Health*, vol. 21, pp. 1–7, 2021. doi: https://doi.org/10.1186/s12903-021-01513-3.

[10] H. Li, Y. Xu, Y. Lei, Q. Wang, and X. Gao, "AUTOMATIC CLASSIFICATION FOR SAGITTAL CRANIOFACIAL PATTERNS BASED ON DIFFERENT CONVOLUTIONAL NEURAL NETWORKS," *Diagnostics*, vol. 12, no. 6, p. 1359, 2022. doi: https://doi.org/10.3390/diagnostics12061359.

[11] J. N. Zhang *et al.*, "DEEP LEARNING-BASED PREDICTION OF MANDIBULAR GROWTH TREND IN CHILDREN WITH ANTERIOR CROSSBITE USING CEPHALOMETRIC RADIOGRAPHS," *BMC Oral Health*, vol. 23, no. 1, Dec. 2023, doi: https://doi.org/10.1186/s12903-023-02734-4.

[12] P. Tiwari *et al.*, "CNN BASED MULTICLASS BRAIN TUMOR DETECTION USING MEDICAL IMAGING," *Comput Intell Neurosci*, vol. 2022, 2022. doi: https://doi.org/10.1155/2022/1830010.

[13] N. Kumar, T. Lakshmi, D. Slavakkam, and R. Ch, "INTEGRATED PREDICTIVE ANALYSIS FOR PERIODONTAL DISEASE AND MALOCCLUSION DETECTION IN DENTISTRY USING DEEP LEARNING AND CNN-BASED DECISION MAKING," 2023, doi: 10.21203/rs.3.rs-3177552/v1.

[14] G. Celik, "DETECTION OF COVID-19 AND OTHER PNEUMONIA CASES FROM CT AND X-RAY CHEST IMAGES USING DEEP LEARNING BASED ON FEATURE REUSE RESIDUAL BLOCK AND DEPTHWISE DILATED CONVOLUTIONS NEURAL NETWORK," *Appl Soft Comput*, vol. 133, p. 109906, 2023. doi: https://doi.org/10.1016/j.asoc.2022.109906.

[15] S. Aksoy, B. Kiliç, and T. Süzek, "COMPARATIVE ANALYSIS OF THREE MACHINE LEARNING MODELS FOR EARLY PREDICTION OF SKELETAL CLASS-III MALOCCLUSION FROM PROFILE PHOTOS," *Mugla Journal of Science and Technology*, vol. 8, no. 2, pp. 22–30, Dec. 2022, doi: https://doi.org/10.22531/muglajsci.1108397.

[16] G. S. Demircan, B. Kılıç, and T. Önal-Süzek, "EARLY DIAGNOSIS AND PREDICTION OF SKELETAL CLASS III MALOCCLUSION FROM PROFILE PHOTOS USING ARTIFICIAL INTELLIGENCe," in *8th European Medical and Biological Engineering Conference: Proceedings of the EMBEC 2020, November 29–December 3, 2020 Portorož, Slovenia*, Springer, 2021, pp. 434–448. doi: https://doi.org/10.1007/978-3-030-64610-3_50.

[17] L. Nan *et al.*, "AUTOMATED SAGITTAL SKELETAL CLASSIFICATION OF CHILDREN BASED ON DEEP LEARNING," *Diagnostics*, vol. 13, no. 10, p. 1719, 2023. doi: https://doi.org/10.3390/diagnostics13101719.

[18] G. Li *et al.*, "PRACTICES AND APPLICATIONS OF CONVOLUTIONAL NEURAL NETWORK-BASED COMPUTER VISION SYSTEMS IN ANIMAL FARMING: A REVIEW," *Sensors*, vol. 21, no. 4, p. 1492, 2021. doi: https://doi.org/10.3390/s21041492.

[19] S. Y. Chaganti, I. Nanda, K. R. Pandi, T. G. Prudhvith, and N. Kumar, "IMAGE CLASSIFICATION USING SVM AND CNN," in *2020 International conference on computer science, engineering and applications (ICCSEA)*, IEEE, 2020, pp. 1–5. doi: https://doi.org/10.1109/ICCSEA49143.2020.9132851.

[20] O. O. Oladimeji and A. O. J. Ibitoye, "BRAIN TUMOR CLASSIFICATION USING RESNET50-CONVOLUTIONAL BLOCK ATTENTION MODULE," *Applied Computing and Informatics*, no. ahead-of-print, 2023. doi: https://doi.org/10.1108/ACI-09-2023-0022.