






## Original Article

# Automated cervical vertebral maturation staging using deep learning: Enhancing accuracy through random oversampling and memory optimization

Noraina Hafizan Norman<sup>1</sup> , Marshima Mohd Rosli<sup>2</sup> , Nagham Mohammed Al-Jaf<sup>3</sup> , Norhasmira Mohammad<sup>4</sup> , Mohd Yusmialdil Putera Mohd Yusof<sup>5</sup> 

<sup>1</sup>Centre for Paediatric Dentistry and Orthodontic Studies, Faculty of Dentistry, Universiti Teknologi MARA, Sungai Buloh, <sup>2</sup>Department of Computer Science, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, <sup>3</sup>Centre for Comprehensive Care Studies, Faculty of Dentistry, Universiti Teknologi MARA, Sungai Buloh, <sup>4</sup>Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, <sup>5</sup>Institute of Pathology, Laboratory and Forensic Medicine, Universiti Teknologi MARA, Sungai Buloh, Malaysia.



### \*Corresponding author:

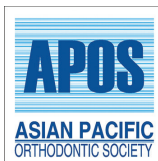
Dr. Mohd Yusmialdil Putera Mohd Yusof,  
Professor, Institute of Pathology, Laboratory and Forensic Medicine (I-PPerForM), Universiti Teknologi MARA, Jalan Hospital, Sungai Buloh, Malaysia.

yusmialdil@uitm.edu.my

Received: 10 December 2024  
Accepted: 21 March 2025  
Epub Ahead of Print: 26 June 2025  
Published:

DOI  
10.25259/APOS\_322\_2024

### Quick Response Code:



## ABSTRACT

**Objectives:** This study introduces a customized deep convolutional neural network (DCNN) framework for automated classification of cervical vertebral maturation stages (CVMS) from lateral cephalometric radiographs, with targeted strategies to address class imbalance and training inefficiencies.

**Material and Methods:** A total of 922 radiographs from subjects aged 7–20 years were independently assessed for CVMS by two orthodontists. Images meeting quality criteria were preprocessed to isolate C2–C4 cervical vertebrae. To address the class imbalance, random oversampling (ROS) was applied. The dataset was split into 70% training, 30% validation, and an additional 10% unseen test set to evaluate model generalization. A custom DCNN model was developed with hyperparameters tuning through random search and trained using Adam optimizer and categorical cross-entropy loss. Early stopping was implemented to prevent overfitting and ensure optimal model convergence. In addition, a memory reset function was applied before each training session to release memory and reset the model's weights, optimizing memory usage and preventing any unwanted bias accumulation during the training process.

**Results:** Initially, the model showed high training accuracy (98%) but but poor generalization (57% validation accuracy) due to dataset imbalance. After applying ROS, dataset restructuring, and early stopping, the model's validation accuracy improved to 88%. On unseen data, the model achieved 76% accuracy, demonstrating better generalization. The recall analysis revealed significant underestimation for CVMS 4 and CVMS 5 (21% and 15% misclassifications), while CVMS 1 and CVMS 6 exhibited minimal misclassifications (8%), mainly within adjacent stages, indicating reasonable stage progression accuracy.

**Conclusion:** This study highlights the potential of a fully automated DCNN for CVMS classification with promising results. Future work will focus on enhancing stage differentiation, improving classification accuracy, and leveraging advanced AI techniques to enhance model robustness and generalization.

**Keywords:** Cervical vertebral maturation, Class imbalance, Deep convolutional neural network, Hyperparameter optimization, Lateral cephalometric radiographs, Random oversampling

## INTRODUCTION

Accurate assessment of skeletal maturity is crucial for successful orthodontic treatment.<sup>[1,2]</sup> Precisely determining the timing of accelerated growth and skeletal development is essential

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, transform, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

©2025 Published by Scientific Scholar on behalf of APOS Trends in Orthodontics

for optimizing treatment outcomes and minimizing the need for complex surgical interventions.<sup>[3-5]</sup> Beyond orthodontics, bone age assessment is valuable for pediatric and forensic medicine such as determining developmental stage, final height, and legal age in cases where identification is missing.<sup>[6,7]</sup> This information aids in diagnosing growth disorders, planning treatments, and forensic investigations.<sup>[8]</sup>

Radiographic analysis is commonly employed to assess skeletal maturation, pubertal development, and growth potential.<sup>[9,10]</sup> Conventionally, hand-wrist radiographs have served as the gold standard for evaluating skeletal age, offering a standardized method for comparison.<sup>[6,11,12]</sup> However, this technique necessitates specialized interpretation skills and exposes patients to ionizing radiation.<sup>[10,11,13,14]</sup> Alternatively, cervical vertebral maturation (CVM) evaluates skeletal maturity by analyzing morphological changes in cervical vertebrae on lateral cephalometric radiographs, a routine orthodontic diagnostic image.<sup>[15]</sup> Although CVM correlates well with hand-wrist assessments and avoids additional radiation, its application can be complex and time-consuming.<sup>[16]</sup> Despite these challenges, CVM offers a potential advantage in orthodontic diagnosis by providing a non-invasive method for evaluating skeletal growth and development.<sup>[17]</sup>

The modified Baccetti's CVM classification provides a framework for assessing skeletal maturity based on observable changes in cervical vertebral morphology.<sup>[18]</sup> The six-stage CVM model delineates the morphological changes of C2, C3, and C4 vertebrae throughout adolescence. Stage 1 (Initiation) is characterized by trapezoidal vertebral bodies with flat inferior borders. As growth accelerates (Stage 2), concavity develops on the inferior borders of C2 and C3, with vertebral bodies transitioning to rectangular shapes. Stage 3 (Transition) marks a period of rapid morphological change, with increasing concavity and persistent rectangular shapes. Growth deceleration (Stage 4) is associated with pronounced concavity and the emergence of square-shaped vertebral bodies. Stage 5 (Maturation) signifies skeletal maturity with maximal concavity and square vertebral bodies. Finally, Stage 6 (Completion) represents the cessation of growth, characterized by deepened concavity and potentially vertically elongated vertebral bodies.<sup>[19,20]</sup>

Continued research is imperative to refine CVM methodology, establish standardized assessment criteria, and integrate it with other diagnostic tools.<sup>[21]</sup> Artificial intelligence (AI), specifically machine learning (ML), has revolutionized medical image analysis. Deep learning (DL), a subset of ML, employs multi-layered neural networks to learn complex patterns directly from data.<sup>[22,23]</sup> Convolutional neural networks (CNNs), a type of DL architecture, have excelled in image classification tasks, including medical imaging applications like convolution computation, and

backpropagation algorithms which can improve disease diagnosis and forensic analysis.<sup>[24]</sup>

Early studies employed traditional ML algorithms to classify CVM stages.<sup>[25]</sup> Most of these studies have compared semiautomated systems to identify landmarks and analyze the CVM stages.<sup>[13,26,27]</sup> However, recent investigations have increasingly adopted DL techniques, particularly CNNs, due to their superior performance in image analysis.<sup>[28]</sup> While some studies have compared different CNN architectures for CVM classification, most have utilized existing models such as ResNet and Inception, with limited exploration of newer models.<sup>[15]</sup> Newer advances in the AI field have explored fully automated systems to eliminate human landmark identification which may be prone to internal and external errors.<sup>[29]</sup> These studies either used a very complex deep architectural structure or did not develop a new model. If older models were too deep, it might have increased computational cost without improving accuracy. Previous models may have lacked data augmentation techniques such as rotation, flipping, or contrast normalization, leading to overfitting.<sup>[30]</sup> Imbalanced datasets were common due to certain CVM stages being underrepresented, which affected classification accuracy.<sup>[31,32]</sup> Nogueira *et al.*<sup>[15]</sup> recently compared 4 different CNN models (AlexNet,16 VGG16,17 ResNet18,18, and Inception-v3.19) and Shoari *et al.*<sup>[29]</sup> compared a custom model to ResNet 18 for CVM analysis. Both studies found that the model's performance could be enhanced through more extensive data augmentation to improve robustness. The subtle differences between CVM stages, coupled with low image quality and imbalanced dataset distribution, posed challenges. Incorporating additional preprocessing layers, expanding the patient age range, and utilizing expert-labeled data could further optimize the model. In addition, exploring different CNN architectures and hyperparameter optimization techniques may yield improved results.

In real-world clinical applications, DL models frequently encounter challenges related to dataset imbalance, where certain classifications naturally occur more often than others. This imbalance can result in biased predictions, favoring dominant classes while underperforming underrepresented ones. To address this, random oversampling (ROS) was applied to enhance class balance, ensuring more effective learning across all categories.<sup>[33]</sup> In addition, early stopping was implemented to prevent overfitting, a common issue in medical AI models trained on limited datasets. A refined training-validation split and a dedicated unseen test set further improved model generalization. These enhancements contribute to a more robust and adaptable approach, not only for CVM classification but also for broader clinical applications where handling class imbalance and optimizing model reliability are essential.<sup>[31]</sup>

This study aims to address this gap by proposing a new CNN model for automated CVM stage classification on lateral cephalometric radiographs and evaluating its performance rate in detecting CVM processes.

## MATERIAL AND METHODS

A sum of 922 digital lateral cephalometric radiographs was acquired from individuals aged between 7 and 20 years for the purpose of pre-orthodontic assessment and treatment planning. Archived radiographs were obtained for research purposes between October 1, 2023, and April 1, 2024, from Radiology Unit, Faculty of Dentistry, Universiti Teknologi MARA (UiTM). Ethical approval was obtained by the UiTM Research Ethics Committee reference number REC/06/2023 (PG/MR/205) which waived the requirement for informed consent due to the retrospective nature of the study. Data management and analysis were conducted in accordance with the principles outlined in the International Council for Harmonisation Good Clinical Practice Guidelines, Malaysian Good Clinical Practice Guidelines, and the Declaration of Helsinki.

The chronological age of the subjects was determined by subtracting their birth date from the date the radiographs were captured. Only radiograph images devoid of artifacts and distortions, and with clear visibility of the C2, C3, and C4 vertebrae, were considered for inclusion in the investigation. All lateral cephalometric radiographs used in the research were obtained using the X-ray unit with a standardized protocol (73 kVp, 15 mA, and 14.9 s exposure time) adhering to the manufacturer's guidelines for positioning and irradiation. The analysis of the images was conducted on a 24-inch medical display monitor (Philips, Luchu Hsiang, Taiwan) equipped with an NVIDIA QUADROFX 380 graphics card to ensure an optimal visual representation.

The evaluation of CVM was independently conducted by two orthodontists (NHN and NA), each with more than 12 years of research and orthodontic clinical experience. Before grading, the orthodontists underwent training and calibration to ensure consistency. The inter-observer agreement was assessed using the kappa coefficient. To minimize fatigue and maintain accuracy, cephalometric images were assessed in multiple sessions. In cases of image labeling discrepancies among observers, a consensus was reached through re-examination for a final decision.

The methodology for building, training, and evaluating the custom CVM stage (CVMS) classification model is detailed as follows:<sup>[34]</sup>

### Data preprocessing

The dataset used in this study comprises images categorized into six distinct classes. To mitigate class imbalance,

ROS was employed to increase the number of samples in underrepresented classes, ensuring a more uniform distribution across all categories. Class imbalance, where certain classes contain significantly fewer samples, can lead to biased model training, making the model overly sensitive to majority classes while reducing its ability to recognize minority classes. To prevent distribution skew, where oversampling unintentionally creates a new majority class, the number of generated samples was capped at the original majority class size. This strategy preserved dataset balance without introducing artificial bias or overfitting risks.

To standardize input dimensions and enhance computational efficiency, all images were resized to  $128 \times 128$  pixels and processed in batches of 32 during training and validation. Data preprocessing and model training were conducted on a Dell Precision 5690 workstation equipped with an Intel Core Ultra 9 185H processor (2.50 GHz), 32 GB of RAM, and an NVIDIA QUADROFX 380 GPU and a 64-bit x64-based architecture, ensuring robust computational performance. The preprocessing pipeline integrated TensorFlow/Keras for image augmentation, OpenCV for resizing and manipulation, NumPy for numerical operations, and Pandas and Matplotlib for data analysis and visualization. This combination of preprocessing techniques, high-performance hardware, and optimized software facilitated an efficient, scalable, and reproducible workflow, ultimately improving model generalization and computational efficiency.

### Custom model building with ROS

Before finalizing the custom model architecture, we performed hyperparameter tuning using a random search to identify the optimal configuration for our image classification task. Hyperparameter tuning involves experimenting with different values for various model parameters to enhance performance and achieve better results. The model was compiled using the Adam optimizer, known for its adaptive learning rate capabilities, making it well-suited for image classification tasks. The loss function used was categorical cross-entropy, appropriate for multi-class classification, and accuracy was chosen as the evaluation metric.<sup>[35]</sup>

### Hyperparameter tuning with random search

Random search was used to optimize key hyperparameters by randomly sampling values from predefined ranges, allowing for a more efficient exploration of potential configurations compared to grid search. The tuning process focused on critical parameters, including the learning rate, batch size, number of units in dense layers, dropout rate, and convolutional layer parameters such as the number of filters and kernel size. The learning rate, which controlled the step size for updating model weights, was varied between 0.0001

and 0.1 to balance convergence speed and stability. The batch size, determining how many training samples were processed before updating the model's parameters, was tested in the range of 16–128 to identify an optimal trade-off between training speed and generalization. The number of neurons in the fully connected dense layers varied between 64 and 512, influencing the model's ability to learn complex patterns.

To prevent overfitting, the dropout rate was adjusted between 0.2 and 0.5, helping to regulate model complexity by randomly deactivating a fraction of neurons during training. In addition, the convolutional layers, responsible for feature extraction, were optimized by varying the number of filters from 32 to 256 and testing kernel sizes of  $3 \times 3$  and  $5 \times 5$  to determine the most effective spatial feature extraction. A search space was defined for each hyperparameter, and random sampling was used to evaluate different configurations. Each selected combination was used to train the model, and performance was assessed on a validation set. The final model architecture consisted of four convolutional layers with optimized filter sizes, followed by batch normalization, max-pooling layers, and fully connected dense layers with an optimal neuron count and dropout rate. The best-performing configuration was selected based on validation accuracy and loss trends, ensuring improved generalization and robustness in CVMS classification. After assessing the performance metrics, the combination of hyperparameters that produced the best results was selected.<sup>[36]</sup>

### Final model architecture

Based on the results of the hyperparameter search, the final model architecture is designed as follows: The model starts with an input layer that processes only the cropped cervical vertebrae (CV2, CV3, and CV4), ensuring the AI focuses exclusively on CVM-relevant regions. Each input image is resized to  $128 \times 128$  pixels with three color channels (RGB) to maintain uniformity and optimize computational efficiency. To enhance contrast and improve feature extraction, normalization techniques are applied, followed by data augmentation (including rotation, flipping, and brightness adjustments) to improve model generalization.

The feature extraction process begins with a convolutional layer containing 96 filters and a kernel size of 3, which captures initial patterns in the vertebral structures. This is followed by a second convolutional layer with 112 filters and a kernel size of 3, further refining extracted features. To reduce spatial dimensions and minimize parameters, a global average pooling layer aggregates information from the feature maps. A fully connected dense layer with 112 units and Rectified Linear Uni (ReLU) activation follows, introducing non-linearity to enhance the model's learning capacity. To mitigate overfitting, a dropout layer (rate = 0.2) randomly deactivates a fraction of units during training.

Finally, the model concludes with an output layer of six neurons, each corresponding to one of the six CVMSs, with a softmax activation function generating probability distributions across the classes. By integrating region-specific preprocessing, data augmentation, and an optimized DL architecture, the model ensures precise feature extraction, robust learning, and accurate CVMS classification.<sup>[33]</sup>

### Training the custom model

The custom model was trained on the training dataset using the fit method, which performs backpropagation to iteratively update the model's weights over multiple epochs. For this study, a custom DCNN model was designed, with hyperparameters optimized using random search, and trained for 100 epochs with the Adam optimizer and categorical cross-entropy loss. Early stopping is a widely used regularization technique in DL that prevents overfitting by halting training once the model's performance stops improving. It works by monitoring a specific metric, such as validation loss or validation accuracy, and stopping training if no significant improvement is observed after a predefined number of epochs. For instance, if validation loss does not decrease for five consecutive epochs, training is terminated, and the model reverts to the best-performing weights to ensure optimal generalization. In addition, a minimum improvement threshold (min delta) can be set to avoid stopping too early due to minor fluctuations in performance. This approach helps maintain model efficiency by reducing unnecessary training cycles while ensuring that the model does not memorize noise from the training data. By applying early stopping, DL models, especially in medical imaging applications, can achieve better generalization and robustness, particularly when trained on limited or imbalanced datasets. During each epoch, the training dataset was divided into mini-batches, and the model was exposed to these mini-batches sequentially, allowing it to learn and update weights gradually.

After each epoch, the model's performance was evaluated on a separate validation dataset. This evaluation helps monitor for overfitting and ensures that the model's performance generalizes well to unseen data. Performance metrics such as training and validation accuracy and loss for each epoch were recorded, providing insights into the model's learning progress and effectiveness.

This recorded training history was used for detailed analysis and to make any necessary adjustments to improve the model's performance. The validation dataset results guided decisions to refine the model and ensure that it was learning effectively without overfitting.<sup>[37]</sup>

### Refined training methodology and data preprocessing strategies

Further improvements were implemented to refine the model's training methodology and enhance classification performance. One of the key modifications was adjusting the training validation split from 80-20 to 70-30, providing a larger validation set for a more reliable performance assessment. This adjustment allowed the model to generalize better and reduced the risk of overfitting by exposing it to a more diverse set of validation samples.

The ROS was applied before training rather than during or after data processing. This ensured that all CVMS stages had sufficient representation in the training data, allowing the model to learn distinctive features for each stage more effectively. Another critical strategy involved resetting the model's weights and clearing the computational graph before each training session. A memory reset function was applied after ROS and before model training to optimize resource management. This approach reduced memory accumulation, improved training stability, and enhanced overall classification performance using `K.clear_session()` in Keras. By invoking this function, each training cycle started fresh, preventing unwanted bias accumulation from previous runs. This technique ensured that the model effectively learned from the newly structured dataset, enhancing its stability and generalization.

The training methodology was further refined through optimized hyperparameter tuning, focusing on adjusting the learning rate, batch size, dropout rate, and the number of units in dense layers. Early stopping was also employed to prevent overfitting and halting training when validation loss plateaued. These refinements contributed to a more stable and efficient training process, ensuring that the model learned meaningful patterns for accurate classification.

### Evaluation

Evaluation metrics included loss and accuracy, with validation accuracy indicating how well the model performed on data not used during training. To gain a deeper understanding of the model's performance across different classes, the confusion matrix was computed. This matrix provides detailed insights into the number of true positive, false positive, false negative, and true negative predictions for each class, revealing any specific classes where the model might be struggling. The predicted labels for the validation set were obtained using the `predict` method, and these predictions were compared with the true labels to construct the confusion matrix.<sup>[38]</sup>

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics 23.0. Inter-rater reliability was assessed with the kappa coefficient, with values

interpreted as follows: Slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–0.99) agreement.

### RESULTS

The Kappa coefficient of 0.87 indicates an almost perfect level of agreement between the two observers in determining CVMSs, suggesting high reliability and consistency in the data. [Table 1] presents the demographic breakdown of the patient population, along with the distribution of CVMSs identified through visual assessment.

Initially, the custom model was developed without implementing ROS. During training, the model achieved a perfect accuracy of 100%. However, this high accuracy did not translate to validation performance, which was only 57%. This discrepancy suggests that while the model fits the training data exceptionally well, it struggled with generalization, likely due to class imbalance or insufficient representation of minority classes. [Figure 1] shows learning curves of training and validation accuracy without applying the ROS.

To address the class imbalance, ROS was applied. Initially, the dataset had varying numbers of images per class, ranging from 32 images for class CVMS 1 to 296 images for class CVMS 5. ROS increased the number of images in each class to 296, resulting in a final dataset of 1420 images for training and 356 images for validation. This increase in dataset size explains the discrepancy between the image count mentioned in different sections of the study. While the dataset originally contained fewer than 1000 images, the application of ROS expanded it to over 1700 images for training and validation, ensuring a balanced distribution across all six classes. [Figure 2] illustrates the confusion matrix for the custom model before applying ROS, where significant misclassifications were observed, particularly among classes CVMS 2 through CVMS 6.

The training process does not start from scratch for each stage; instead, the model is trained in a single end-to-end

**Table 1:** Descriptive statistics of the patient's age and cervical vertebral maturation stage.

CVM stages	Mean age (Years)±SD	n (%)
CVM Stage 1	7.59±1.64	32 (3.47)
CVM Stage 2	9.75±1.53	47 (5.10)
CVM Stage 3	10.61±1.24	97 (10.52)
CVM Stage 4	12.16±1.31	196 (21.26)
CVM Stage 5	13.62±1.35	296 (32.10)
CVM Stage 6	17.21±3.04	254 (27.55)
Total	11.82±3.35	922 (100)

CVM: Cervical vertebral maturation, SD: Standard deviation

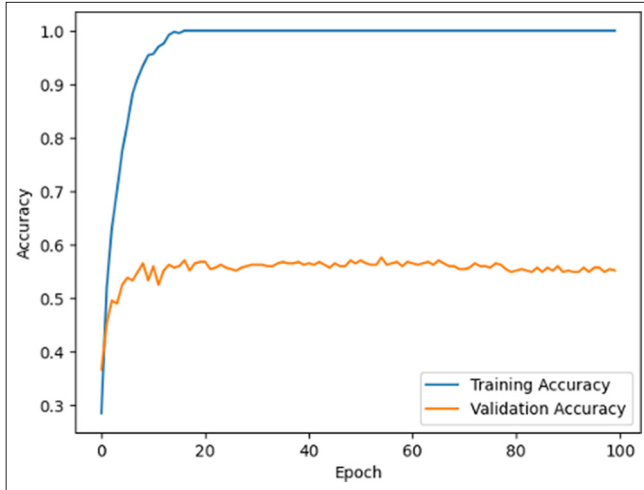


Figure 1: Learning curves of training and validation accuracy without random oversampling.

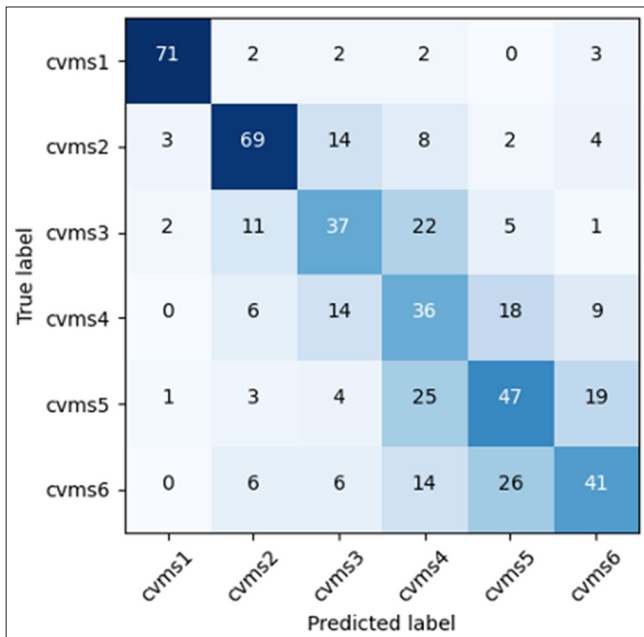


Figure 2: Confusion matrix for the custom model without random oversampling. cvms: Cervical vertebral maturation stage.

session, learning to classify all six CVMSs simultaneously. Before each training run, previous training states are erased by reinitializing the model’s weights and reloading the dataset. Specifically, this is achieved by resetting the model’s weights, clearing the computational graph using `K.clear_session()` in Keras, and reloading the balanced dataset before retraining. This ensures that each training session begins without residual knowledge from previous runs, allowing the model to learn from the newly structured dataset effectively.

The hyperparameter search has been completed, yielding the optimal configurations for the model. [Figure 3] presents

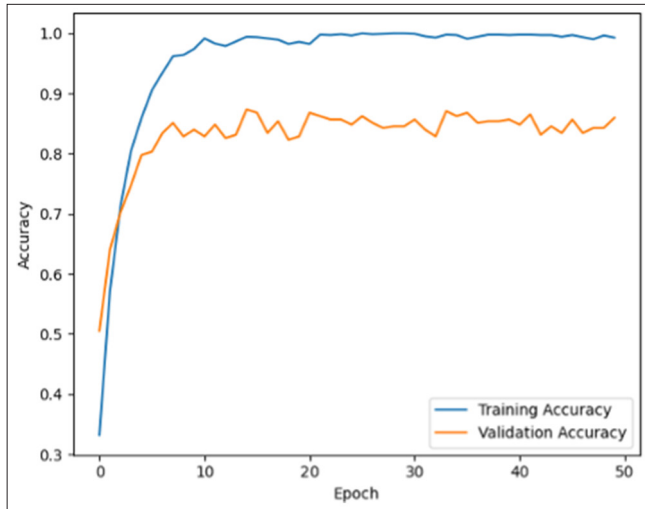
Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 126, 126, 96)	2688
max_pooling2d_2 (MaxPooling 2D)	(None, 63, 63, 96)	0
conv2d_3 (Conv2D)	(None, 61, 61, 112)	96880
max_pooling2d_3 (MaxPooling 2D)	(None, 30, 30, 112)	0
flatten_1 (Flatten)	(None, 100800)	0
dense_2 (Dense)	(None, 112)	11289712
dropout_1 (Dropout)	(None, 112)	0
dense_3 (Dense)	(None, 6)	678
-----		
Total params: 11,389,958		
Trainable params: 11,389,958		
Non-trainable params: 0		

Figure 3: Detailed layer configuration of the custom model.

the optimal hyperparameters identified through the search. For the first convolutional layer, the optimal number of filters was determined to be 96, with a kernel size of 3. This configuration was selected to effectively capture initial features from the input images. The second convolutional layer was optimized with 112 filters and the same kernel size of 3, enhancing the model’s ability to refine and extract more complex features.

In the fully connected dense layer, the optimal number of units was found to be 112. This setting helps the model learn intricate patterns and relationships in the data. To mitigate overfitting, a dropout rate of 0.2 was identified as optimal, providing a balance between regularization and model capacity. These hyperparameters were chosen to maximize the model’s performance and efficiency in handling the image classification task, ensuring robust feature extraction and effective learning.

During the training of the model over 100 epochs, significant improvements in performance were observed. Initially, the model achieved a loss of 1.7053 and an accuracy of 33.24% on the training set, with a validation accuracy of 50.56%. By the fifth epoch, training accuracy had increased to 85.92%, with validation accuracy rising to 79.78%. As training progressed, accuracy continued to improve, reaching 97.39% by epoch 10, although validation accuracy experienced fluctuations, peaking at 87.36% in epoch 15. After the 50<sup>th</sup> epoch, training was stopped due to the early stopping function. Despite some variability in validation accuracy, ranging between 82.30% and 87.36%, the model achieved a final accuracy of 85.96% by epoch 50, indicating a robust model with good generalization capabilities. [Figure 4] shows the learning curve for classifying CVMS, illustrating the model’s performance.



**Figure 4:** Learning curves of training and validation accuracy with random oversampling.

The results presented reflect the model's performance after testing on the validation dataset. [Table 2] illustrates the classification report obtained after the model evaluation. For CVMS 1, the model achieved perfect accuracy and recall, indicating flawless identification of this class with very few false positives or false negatives. The CVMS 2 also performed exceptionally well, with a high precision of 95.8% and a recall of 98.6%, suggesting strong predictive capability and reliable classification. The CVMS 3 demonstrated balanced performance with a precision and recall of 96.0%, reflecting an effective and consistent ability to identify this class.

However, CVMS 4 showed slightly lower precision at 85.7% and recall at 87.5%, pointing to some challenges in minimizing false positives and negatives. The performance of CVMS 5 was notably weaker, with a lower precision of 77.0% and recall of 71.2%, indicating difficulties in accurately classifying this class, which may be due to complex feature differentiation. The CVMS 6 had moderate performance with precision and recall around 79.0% and 80.0%, respectively, suggesting reasonable effectiveness but room for improvement.

The overall classification accuracy of the model is 88.2%, reflecting a generally strong performance. The macro and weighted averages further confirm that the model is performing well across different classes, with the macro average indicating balanced performance across all classes and the weighted average accounting for class imbalances. The classification report highlights that while the model excels in several classes, attention should be given to improving the classification of CVMS 5 to enhance overall robustness. The confusion matrix in [Figure 5] reveals varied performance across classes. This study focuses on classifying CVMSs rather than analyzing the shape and outline of

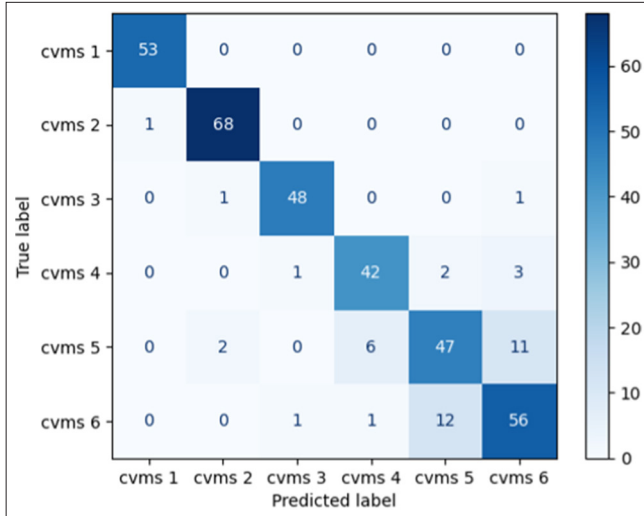
**Table 2:** The precision, recall, and F1-score values calculated according to the confusion matrix.

Stage	Precision	Recall	F1-score
1	0.98	1.00	0.99
2	0.96	0.99	0.97
3	0.96	0.96	0.96
4	0.85	0.88	0.87
5	0.77	0.71	0.74
6	0.79	0.80	0.79
Accuracy	0.88	0.88	0.88

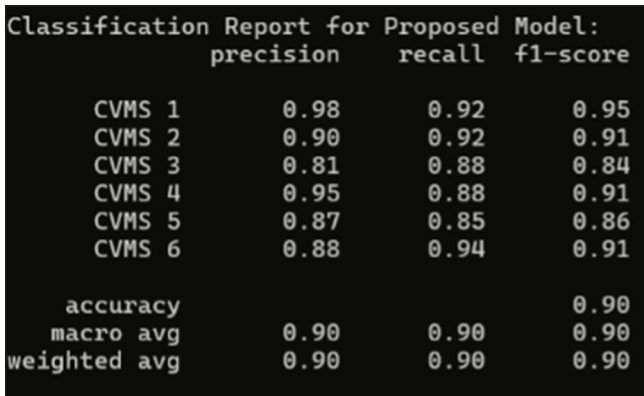
individual vertebrae. Since the regions of interest (ROI) were pre-cropped to include only the relevant cervical vertebrae, the model is trained specifically on these maturation stages, eliminating the need for separate vertebra-level analysis.

Further improvements were achieved by refining the training approach, adjusting the training-validation split to 70–30, and applying ROS before training. These modifications led to a significant performance boost, with the model achieving an overall accuracy of 90%. [Figure 6] shows the classification report for the proposed model. The macro and weighted average F1-scores of 0.90 indicate strong classification consistency across all CVMS stages. One of the key improvements was the balancing of class distribution using ROS. This technique ensured that all CVMS stages had sufficient representation, leading to a noticeable improvement in recall, particularly for CVMS 5 (0.85 vs. previous 0.71). This directly addressed previous misclassification issues, where CVMS 5 was frequently confused with adjacent stages due to an insufficient number of training samples.

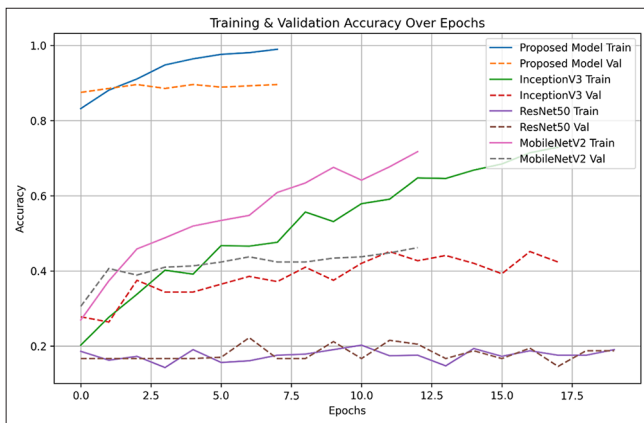
In addition, `K.clear_session()` was applied before each training session to reset the model's weights and clear the computational graph. This ensured that every training cycle started fresh, preventing unwanted bias accumulation and allowing the model to learn effectively from the newly structured dataset. The refined training strategy also contributed to higher precision and recall in later CVMS stages, which previously exhibited performance inconsistencies. CVMS 4's precision improved to 0.95, reducing false positives, while CVMS 6's recall increased to 0.94, indicating better sensitivity in detecting skeletal maturity progression. These enhancements resulted in a more robust and stable model capable of distinguishing all CVMS stages with improved reliability. By integrating these refinements, the model demonstrated greater classification robustness and stability, particularly in distinguishing the more challenging CVMS 5 stage. The applied techniques further reinforce the model's adaptability for other medical classification tasks, highlighting its potential real-world clinical applicability beyond CVM staging.



**Figure 5:** Confusion matrix for custom model with random oversampling. cvms: Cervical vertebral maturation stage.



**Figure 6:** Improved classification report for proposed model. CVMS: Cervical vertebral maturation stage.



**Figure 7:** Accuracy comparison of the proposed model versus pre-trained models (InceptionV3, ResNet50, and MobileNetV2) under identical experimental settings.

The accuracy comparison graph, presented in [Figure 7], illustrates the performance of the proposed method against several pre-trained models, including InceptionV3, ResNet50, and MobileNetV2, under identical experimental settings. The same training parameters, dataset preprocessing, and hyperparameter configurations were applied to ensure a fair comparison across all models. The proposed model achieves the highest training accuracy, approaching 100%, while maintaining a stable validation accuracy above 85%. The minimal gap between the training and validation curves indicates strong generalization, suggesting that the model effectively learns the classification patterns without significant overfitting.

Among the pre-trained models, InceptionV3 exhibits lower training accuracy, peaking around 40–50%, with fluctuating validation accuracy, indicating unstable learning. ResNet50 performs the worst, struggling to surpass 20% in both training and validation accuracy, suggesting that it fails to learn effectively under the given settings. MobileNetV2 shows moderate performance, with training accuracy steadily increasing beyond 60%. However, its validation accuracy remains inconsistent and lower than the proposed model, suggesting weaker generalization capability.

[Figure 8] presents the classification report for the unseen dataset. The model achieved an overall accuracy of 74%, with macro and weighted average F1 scores of 74% and 79%, respectively. CVMS1 and CVMS3 showed perfect precision (100%) but lower recall (75% and 50%), suggesting underprediction of these classes. CVMS5 had the highest recall (100%) but a slightly lower precision (80%), indicating some misclassifications as CVMS5. CVMS2 performed weakest, with 50% precision and 67% recall, often being confused with other classes. CVMS4 and CVMS6 exhibited moderate performance, with precision and recall between 60% and 75%.

## DISCUSSION

Resizing the training and validation images is essential for several reasons. First, ML models, particularly CNNs, require input images of the same size for the architecture to function correctly. Resizing ensures uniform dimensions, leading to more consistent and reliable training outcomes. In addition, smaller, uniformly sized images reduce computational load and memory usage, making the training process faster and more efficient. Using a standard size like  $128 \times 128$  ensures uniformity across the dataset, which is crucial for model training. It avoids the introduction of bias that could occur if images of varying sizes were used. Besides, smaller image dimensions reduce the computational load and memory usage during training. By resizing to  $128 \times 128$ , the model can process images more quickly, which is particularly

Classification Report:			
	precision	recall	f1-score
CVMS1	1.00	0.75	0.86
CVMS2	0.50	0.67	0.57
CVMS3	1.00	0.50	0.67
CVMS4	0.60	0.75	0.67
CVMS5	0.80	1.00	0.89
CVMS6	0.75	0.75	0.75
accuracy			0.74
macro avg	0.78	0.74	0.73
weighted avg	0.79	0.74	0.74

**Figure 8:** Classification performance on unseen dataset. CVMS: Cervical vertebral maturation stage.

important when working with large datasets or complex models. In clinical applications, if input images differ in size, it may impact model accuracy due to variations in spatial features. While the model was trained with this fixed size, resizing clinical images to  $128 \times 128$  before inference helps maintain accuracy and ensures compatibility with the trained network. Deviations from this resolution may require additional preprocessing steps or adaptive resizing techniques to preserve model performance.

Standardizing image sizes also prevents inconsistencies that could affect the model's learning process. For CNNs, resizing ensures that the model can consistently extract relevant features across all images. Furthermore, resizing to the specific size used by pre-trained models (e.g.,  $224 \times 224$ ) ensures compatibility, allowing for effective transfer learning and leveraging pre-trained weights. Overall, resizing is a crucial pre-processing step that optimizes the dataset for better model performance and efficient use of resources.

Developing a network from scratch without annotations offers flexibility, deep understanding, and optimization opportunities, enabling tailored architectures for specific tasks. It fosters skills enhancement, allows for unsupervised or self-supervised learning, encourages innovative approaches, and provides full control over the data processing and training pipeline, making it ideal for research and experimentation despite requiring significant effort and expertise.

ROS aims to balance the class distribution by generating synthetic samples for minority classes, providing a more representative dataset. This adjustment is intended to improve the model's generalization capabilities and enhance validation performance, thereby reducing the gap between training and validation accuracy and leading to more reliable classification results across all classes. When performing ROS, the classification performance of the majority class can sometimes be quite low due to several factors. First, ROS can lead to overfitting on the minority class as the algorithm might memorize the duplicated instances rather than learn

generalizable patterns, resulting in poor performance on unseen data, especially affecting the majority class. Second, although oversampling balances the class distribution, it does not create new information, and the oversampled data might not represent the underlying distribution of the classes well, causing difficulty in generalization. In addition, oversampling can amplify noise present in the minority class; duplicating noisy instances makes the model more prone to misclassify the majority class. Furthermore, by balancing the classes, the model might shift its focus toward the minority class, leading to better performance for the minority class but potentially lowering the performance of the majority class.

Initially, the model exhibited a high training accuracy of 100%, but validation accuracy was substantially lower at 57%, indicating issues with class imbalance. To address this, ROS was implemented. This technique increased the number of images for each class to balance the dataset, enhancing its representation and improving model training. As a result, the final dataset consisted of 1420 images for training and 356 for validation. The application of ROS led to a notable improvement in the model's performance metrics, addressing the previously observed misclassifications, particularly in classes CVMS 2 through CVMS 6. The confusion matrix and performance metrics post-ROS demonstrated a more balanced and effective classification across all classes, reducing the number of misclassifications and improving overall model robustness.

Li *et al.*<sup>[37]</sup> collected an extensive dataset of 10,200 radiographs, significantly larger than previous studies. They utilized YOLOv3 for detecting ROI and achieved an overall accuracy of 70%. Their approach highlighted the potential of DL in CVM classification and suggested incorporating additional factors, such as intervertebral disc space and dental age, for further improvements. In contrast, our custom network with ROS implementation achieved superior accuracy, with an overall performance of 88%. Although our customized model showed improvements in overall classification, it still faced challenges in accurately identifying CVMSs 5 and 6. These stages are the majority classes in the dataset, and while ROS effectively balances the dataset by oversampling minority classes, it does not modify the intrinsic distribution of the majority classes. As a result, the model may become more adept at identifying underrepresented stages but struggle to distinguish frequently occurring ones, such as CVMSs 5 and 6, which exhibit subtle structural differences. In addition, the complexity of these later stages, where skeletal maturity progresses with finer variations, adds another layer of difficulty. While ROS improves minority class representation, it does not enhance the model's ability to differentiate subtle variations within majority classes. Consequently, the model may still struggle with frequently occurring stages that require more nuanced feature extraction. To

mitigate this issue, alternative techniques such as synthetic data generation, adaptive weighting strategies, or focal loss could be explored. These approaches would enhance feature diversity, allowing the model to learn more discriminative patterns and improve its classification performance for harder-to-distinguish instances.<sup>[39]</sup>

The improvement in the model accuracy to 88% was likely the result of both ROS and hyperparameter tuning, with each addressing different aspects of model performance. ROS corrected class imbalance by increasing the representation of minority classes, ensuring the model could learn effectively from all classes and reducing bias toward the majority class. Meanwhile, hyperparameter tuning optimized the model's configuration, enhancing its ability to converge efficiently and generalize well to unseen data. Together, these techniques worked synergistically to improve overall accuracy, with the relative contribution of each depending on the initial dataset and model characteristics [Figure 9].

The classification metrics provide a detailed overview of the model's performance across different classes [Table 2]. For CVMS 1, the model achieved perfect accuracy and recall, indicating flawless identification of this class with very few false positives or false negatives. CVMS 2 also performed exceptionally well, with a high precision of 95.8% and a recall of 98.6%, suggesting strong predictive capability and reliable classification. The CVMS 3 demonstrated balanced performance with a precision and recall of 96.0%, reflecting an effective and consistent ability to identify this class.

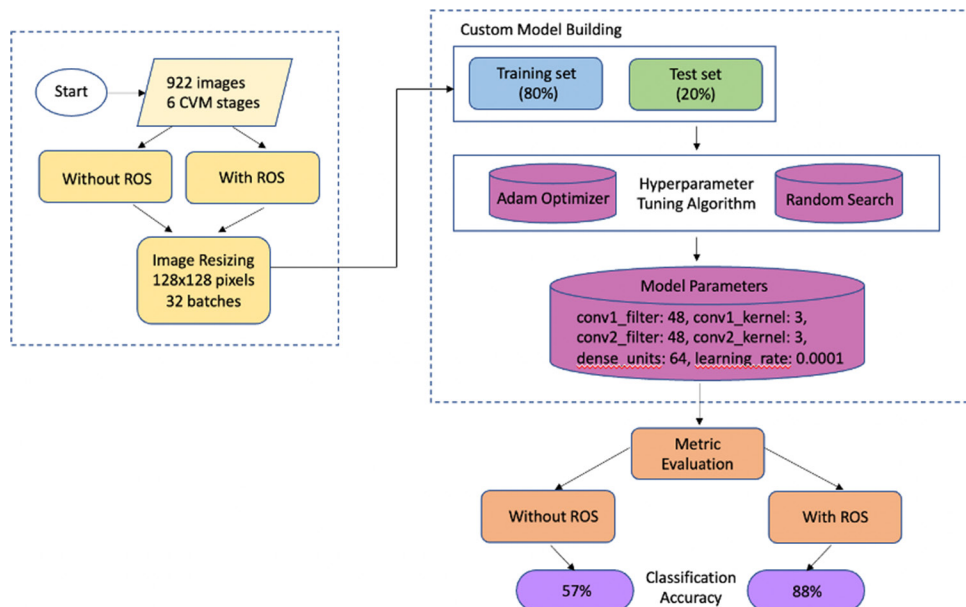
However, CVMS 4 showed slightly lower precision at 85.7% and recall at 87.5%, pointing to some challenges in

minimizing false positives and negatives. The performance of CVMS 5 was notably weaker, with a lower precision of 77.0% and recall of 71.2%, indicating difficulties in accurately classifying this class, which may be due to complex feature differentiation. The CVMS 6 had moderate performance with precision and recall around 79.0% and 80.0%, respectively, suggesting reasonable effectiveness but room for improvement.

Misclassification significantly occurred in CVMS 6 due to the large age gap among the subjects. This age disparity led to high morphological variation within the dataset. As individuals grow, their morphological features can change considerably, especially in developmental stages covered by CVMS 6. These variations make it difficult for classification algorithms to categorize the subjects consistently and accurately. The high degree of morphological diversity introduces complexity in pattern recognition and model training, resulting in inconsistent classification results.<sup>[40]</sup>

A comparison between the 80-20 and 70-30 data assignments highlights significant improvements in classification performance, particularly in recall and stage misclassification rates. Under the previous 80-20 split, the model exhibited a higher tendency to misclassify CVMS stages, struggling with underestimation issues, especially for CVMS 5 and CVMS 4. With the current 70-30 split, where a larger validation set was used, the model's generalization ability improved, leading to better stage differentiation and reduced misclassification rates.

The recall values indicate the percentage of samples misclassified into earlier stages. CVMS 5, which previously



**Figure 9:** Summary of custom deep convolutional neural network model and its performance. CVM: Cervical vertebral maturation, ROS: Random Oversampling.

faced the highest underestimation, saw a notable reduction in misclassified samples, now with 21% misclassified into earlier stages. CVMS 4 followed with 15%, while CVMS 6 and CVMS 1 exhibited an 8% underestimation rate. Importantly, these errors remained within a one-stage difference, meaning that while some misclassifications persist, the model effectively preserves the logical progression of CVMS stages.

The shift to a 70-30 split provided a larger validation set, allowing for more comprehensive performance evaluation and model tuning. This led to a more balanced classification across all categories, reducing extreme cases of misclassification observed in the 80-20 setup. The refined dataset distribution, combined with improved preprocessing techniques such as ROS and systematic weight resetting, contributed to a more stable and reliable classification performance.

To mitigate these issues, alternative techniques or combinations of techniques can be considered. Methods like the Synthetic Minority Over-sampling Technique (SMOTE) generate synthetic instances rather than duplicating existing ones, helping to create more diverse samples for the minority class. Adaptive Synthetic Sampling, a variant of SMOTE, focuses more on generating synthetic samples for minority-class instances that are harder to classify. Ensemble methods, which combine the results of multiple models, can often balance the bias towards any particular class.

Future research should focus on refining the model to further enhance stage differentiation, particularly for closely related stages. Exploring more advanced architectures, incorporating skip connections, or adopting alternative pooling methods could improve feature extraction. In addition, leveraging hyperparameter optimization techniques such as Bayesian Optimization or Genetic Algorithms may enhance classification performance. Integrating explainable AI techniques could also provide greater transparency in model decision-making, facilitating clinical adoption and validation.

## CONCLUSION

This study demonstrates the potential of a fully automated DCNN for CVMS classification, with promising results and areas for refinement, particularly for challenging stages like CVM 5. Future research will focus on enhancing stage differentiation, improving classification accuracy, and leveraging advanced AI techniques to improve model robustness and generalization.

**Ethical approval:** The research/study was approved by the Institutional Review Board at UiTM Research Ethics Committee, number REC/06/2023 (PG/MR/205), dated 15th September 2023.

**Declaration of patient consent:** Patient consent is not required as there are no patients in this study.

**Financial support and sponsorship:** Prototype Research Grant Scheme PRGS/1/2023/SKK07/UITM/02/1.

**Conflicts of interest:** There are no conflicts of interest.

**Use of artificial intelligence (AI)-assisted technology for manuscript preparation:** The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript, and no images were manipulated using AI.

## REFERENCES

- Hagg U, Pancherz H. Dentofacial orthopaedics in relation to chronological age, growth period and skeletal development. An analysis of 72 male patients with class II division 1 malocclusion treated with the Herbst appliance. *Eur J Orthod* 1988;10:169-76.
- Gray S, Bennani H, Kieser JA, Farella M. Morphometric analysis of cervical vertebrae in relation to mandibular growth. *Am J Orthod Dentofacial Orthop* 2016;149:92-8.
- Abdulla Alkhal H, Wong RW, Rabie AB. Correlation between chronological age, cervical vertebral maturation and Fishman's skeletal maturity indicators in Southern Chinese. *Angle Orthod* 2008;78:591-6.
- Perinetti G, Contardo L. Reliability of growth indicators and efficiency of functional treatment for skeletal class II malocclusion: Current evidence and controversies. *Biomed Res Int* 2017;2017:1367691.
- Montasser MA. Craniofacial growth spurt in class I subjects. *Am J Orthod Dentofacial Orthop* 2019;155:473-81.
- Kim DW, Kim J, Kim T, Kim T, Kim YJ, Song IS, *et al.* Prediction of hand-wrist maturation stages based on cervical vertebrae images using artificial intelligence. *Orthod Craniofac Res* 2021;24:68-75.
- Perinetti G, Braga C, Contardo L, Primozic J. Cervical vertebral maturation: Are postpubertal stages attained in all subjects? *Am J Orthod Dentofacial Orthop* 2020;157:305-12.
- Kök H, Izgi MS, Acilar AM. Determination of growth and development periods in orthodontics with artificial neural network. *Orthod Craniofac Res* 2021;24:76-83.
- Houston WJ, Miller JC, Tanner JM. Prediction of the timing of the adolescent growth spurt from ossification events in hand-wrist films. *Br J Orthod* 1979;6:145-52.
- Tekin A, Cesur Aydın K. Comparative determination of skeletal maturity by hand-wrist radiograph, cephalometric radiograph and cone beam computed tomography. *Oral Radiol* 2020;36:327-36.
- Fishman LS. Radiographic evaluation of skeletal maturation. A clinically oriented method based on hand-wrist films. *Angle Orthod* 1982;52:88-112.
- Bowden BD. Epiphysial changes in the hand/wrist area as indicators of adolescent stage. *Aust Orthod J* 1976;4:87-104.
- Makaremi M, Lacaule C, Mohammad-Djafari A. Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography. *Entropy* 2019;21:1222.
- Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd ed. Stanford: Stanford University Press; 1959.
- Nogueira-Reis F, Cascante-Sequeira D, Farias-Gomes A, De Macedo MM, Watanabe RN, Santiago AG, *et al.* Determination

- of the pubertal growth spurt by artificial intelligence analysis of cervical vertebrae maturation in lateral cephalometric radiographs. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;138:306-15.
16. Khajah A, Tadinada A, Allareddy V, Kuo CL, Nanda R, Uribe F. Influence of type of radiograph and levels of experience and training on reproducibility of the cervical vertebral maturation method. *Am J Orthod Dentofacial Orthop* 2020;157:228-39.
  17. Stiehl J, Müller B, Dibbets J. The development of the cervical vertebrae as an indicator of skeletal maturity: Comparison with the classic method of hand-wrist radiograph. *J Orofac Orthop* 2009;70:327-35.
  18. Baccetti T, Franchi L, Mcnamara JA Jr. An improved version of the cervical vertebral maturation (CVM) method for the assessment of mandibular growth. *Angle Orthod* 2002;72:316-23.
  19. Baccetti T, Franchi L, McNamara JA Jr. The cervical vertebral maturation (CVM) method for the assessment of optimal treatment timing in dentofacial orthopedics. *Semin Orthod* 2005;11:119-29.
  20. Baccetti T, Franchi L, De Toffol L, Ghiozzi B, Cozza P. The diagnostic performance of chronologic age in the assessment of skeletal maturity. *Prog Orthod* 2006;7:176-88.
  21. Dipalma G, Inchingolo AD, Inchingolo AM, Piras F, Carpentiere V, Garofoli G, *et al.* Artificial intelligence and its clinical applications in orthodontics: A systematic review. *Diagnostics (Basel)* 2023;13:3677.
  22. Kondody RT, Patil A, Devika G, Jose A, Kumar A, Nair S. Introduction to artificial intelligence and machine learning into orthodontics: A review. *APOS Trends Orthod* 2022;12:214-20.
  23. Khanagar SB, Al-Ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, *et al.* Developments, application, and performance of artificial intelligence in dentistry - A systematic review. *J Dent Sci* 2021;16:508-22.
  24. Franco A, Porto L, Heng D, Murray J, Lygate A, Franco R, *et al.* Diagnostic performance of convolutional neural networks for dental sexual dimorphism. *Sci Rep* 2022;12:17279.
  25. Seo H, Hwang J, Jeong T, Shin J. Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs. *J Clin Med* 2021;10:3591.
  26. Kök H, Acilar AM, İzgi MS. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog Orthod* 2019;20:41.
  27. Amasya H, Yildirim D, Aydogan T, Kemaloglu N, Orhan K. Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: Comparison of machine learning classifier models. *Dentomaxillofac Radiol* 2020;49:20190441.
  28. Atici SF, Ansari R, Allareddy V, Suhaym O, Cetin AE, Elnagar MH. Fully automated determination of the cervical vertebrae maturation stages using deep learning with directional filters. *PLoS One* 2022;17:e0269198.
  29. Shoari SA, Sadrolashrafi SV, Sohrabi A, Afrouzian R, Ebrahimi P, Kouhsoltani M, *et al.* Estimating mandibular growth stage based on cervical vertebral maturation in lateral cephalometric radiographs using artificial intelligence. *Prog Orthod* 2024;25:28.
  30. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019;6:60.
  31. Razali MN, Arbaiy N, Lin PC, Ismail S. Optimizing multiclass classification using convolutional neural networks with class weights and early stopping for imbalanced datasets. *Electronics* 2025;14:705.
  32. Wang Y, Rosli MM, Musa N, Li F. Multi-class imbalanced data classification: A systematic mapping study. *Eng Technol Appl Sci Res* 2024;14:14183-90.
  33. Hellín CJ, Olmedo AA, Valledor A, Gómez J, López-Benítez M, Tayebi A. Unraveling the impact of class imbalance on deep-learning models for medical image classification. *Appl Sci* 2024;14:3419.
  34. Mohammad N, Muad AM, Ahmad R, Yusof MY. Accuracy of advanced deep learning with tensorflow and keras for classifying teeth developmental stages in digital panoramic imaging. *BMC Med Imaging* 2022;22:66.
  35. Mohammad N, Muad AM, Ahmad R, Mohd Yusof MY. Reclassification of Demirjian's mandibular premolars staging for age estimation based on semi-automated segmentation of deep convolutional neural network. *Forensic Imaging* 2021;24:200440.
  36. Akay G, Akcayol MA, Özdem K, Güngör K. Deep convolutional neural network-the evaluation of cervical vertebrae maturation. *Oral Radiol* 2023;39:629-38.
  37. Li H, Li H, Yuan L, Liu C, Xiao S, Liu Z, *et al.* The psc-CVM assessment system: A three-stage type system for CVM assessment based on deep learning. *BMC Oral Health* 2023;23:557-68.
  38. Khazaei M, Mollabashi V, Khotanlou H, Farhadian M. Automatic determination of pubertal growth spurts based on the cervical vertebral maturation staging using deep convolutional neural networks. *J World Fed Orthod* 2023;12:56-63.
  39. Alkhalwaldeh IM, Albalkhi I, Naswhan AJ. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World J Methodol* 2023;13:373-8.
  40. Sadeghi TS, Ourang SA, Sohrabniya F, Sadr S, Shobeiri P, Motamedian SR. Performance of artificial intelligence on cervical vertebral maturation assessment: A systematic review and meta-analysis. *BMC Oral Health* 2025;25:187.

**How to cite this article:** Norman NH, Mohd Rosli M, Mohammed Al-Jaf N, Mohammad N, Mohd Yusof MYP. Automated cervical vertebral maturation staging using deep learning: Enhancing accuracy through random oversampling and memory optimization. *APOS Trends Orthod*. doi: 10.25259/APOS\_322\_2024