Fırat Üniversitesi Deneysel ve Hesaplamalı Mühendislik Dergisi

FUJECE

Transfer Öğrenme Tabanlı Derin Öğrenme Yaklaşımlarıyla Servikal Vertebra Matürasyon Safhalarının Sınıflandırılması ve Kemik Yaşı Değerlendirilmesi

Mazhar KAYAOĞLU[1]*, Abdülkadir ŞENGÜR[2], Sabahattin BOR[3], Seda KOTAN[4]

[1]Enformatik Bölüm Başkanlığı, Bingöl Üniversitesi, Bingöl, Türkiye.
[2]Elektrik Elektronik Mühendisliği, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye.
[3]Klinik Bilimler Bölümü, Diş Hekimliği Fakültesi, İnönü Üniversitesi, Malatya, Türkiye.
[4]Klinik Bilimler Bölümü, Diş Hekimliği Fakültesi, Iğdır Üniversitesi, Iğdır, Türkiye.
[1]mkayaoglu@bingol.edu.tr, [2]ksengur@firat.edu.tr, [3]sabahattin.bor@inonu.edu.tr, [4]dtsedakotan@gmail.com

Geliş Tarihi: 14.03.2025
Kabul Tarihi: 30.05.2025
Düzeltme Tarihi: 22.04.2025
doi: https://doi.org/10.62520/fujece.1657886
Araştırma Makalesi

Alıntı: M. Kayaoğlu, A. Şengür, S. Bor ve S. Kotan, "Transfer öğrenme tabanlı derin öğrenme yaklaşımlarıyla servikal vertebra matürasyon safhalarının sınıflandırılması ve kemik yaşı değerlendirilmesi", Fırat Üni. Deny. ve Hes. Müh. Derg., vol. 4, no 2, pp. 393-405, Haziran 2025.

Öz

Bu çalışmada, büyüme ve gelişimi değerlendirmek amacıyla lateral sefalometrik radyografiler kullanılarak servikal vertebra matürasyon (CVM) evrelerinin otomatik sınıflandırılması gerçekleştirilmiştir. Van Yüzüncü Yıl Üniversitesi Diş Hekimliği Fakültesi Ortodonti Anabilim Dalı tarafından sağlanan toplam 4285 radyografi kullanılmıştır. Uzman hekimler tarafından yapılan detaylı değerlendirmeler sonucunda, tanısal doğruluk ve klinik uygunluk kriterlerini karşılayan 3750 görüntü çalışmaya dâhil edilmiştir. Seçilen görüntüler, altı sınıfa (CVMS 1–6) ayrılarak dengeli bir veri seti oluşturulmuş ve NFNet, ConvNeXt V2, EfficientNet V2 ve DeiT3 modelleri kullanılarak sınıflandırma işlemleri gerçekleştirilmiştir. NFNet modeli, %96 eğitim doğruluğu ve %85,7 test doğruluğu ile en yüksek genel performansı sergilemiştir. %95 eğitim doğruluğu ve %86,9 test doğruluğu elde eden ConvNeXt V2, genelleme açısından en dengeli model olarak öne çıkmıştır. EfficientNet V2, %94 eğitim doğruluğuna ulaşmasına rağmen %80,7 test doğruluğu ile sınırlı bir genelleme kapasitesi göstermiştir. DeiT3 modeli ise %93 eğitim doğruluğu ve %77,6 test doğruluğu ile en düşük genelleme kapasitesine sahip olmuştur. NFNet ve ConvNeXt V2, yüksek doğruluk oranları ve dengeli performansları sayesinde güçlü sınıflandırma adayları olarak öne çıkmıştır. NFNet'in eğitim ve test doğruluğu arasındaki %10,3'lük fark genelleme kapasitesinde bir miktar azalmaya işaret ederken, ConvNeXt V2'nin daha dar olan %8,1'lik farkı daha istikrarlı bir performans göstermiştir. Sonuç olarak, NFNet ve ConvNeXt V2, CVM sınıflandırması için umut vadeden modeller olarak belirlenmiştir. Gelecekteki çalışmalarda, bu modellerin performansını artırmak ve klinik uygulanabilirliklerini güçlendirmek için daha büyük veri setleri kullanılması ve hiperparametre optimizasyonunun gerçekleştirilmesi önerilmektedir.

Anahtar kelimeler: Görüntü sınıflandırma, Transfer öğrenimi, Servikal vertebra matürasyonu

*Yazışılan Yazar

İntihal Kontrol: Evet – Turnitin
Şikayet: fujece@firat.edu.tr

Firat Univ Jour. of Exp. and Comp. Eng., 4(2), 393-405, 2025

Firat University Journal of Experimental and Computational Engineering

FUJECE

Classification of Cervical Vertebral Maturation Stages and Bone Age Assessment Using Transfer Learning–Based Deep-Learning Approaches

Mazhar KAYAOĞLU[1]*, Abdülkadir ŞENGÜR[2], Sabahattin BOR[3], Seda KOTAN[4]

[1]Department of Informatics, Bingöl University, Bingöl, Türkiye.
[2]Department of Electrical and Electronics Engineering, Faculty of Technology, Fırat University, Elazığ, Türkiye.
[3]Department of Clinical Sciences, Faculty of Dentistry, İnönü University, Malatya, Türkiye.
[4]Department of Clinical Sciences, Faculty of Dentistry, Iğdır University, Iğdır, Türkiye.
[1]mkayaoglu@bingol.edu.tr, [2]ksengur@firat.edu.tr, [3]sabahattin.bor@inonu.edu.tr, [4]dtsedakotan@gmail.com

Received: 14.03.2025
Accepted: 30.05.2025
Revision: 22.04.2025
doi: https://doi.org/10.62520/fujece.1657886
Research Article

Citation: M. Kayaoğlu, A. Şengür, S. Bor and S. Kotan, "Classification of cervical vertebral maturation stages and bone age assessment using transfer learning–based deep-learning approaches", Firat Univ. Jour.of Exper. and Comp. Eng., vol. 4, no 2, pp. 393-405, June 2025.

Abstract

In this study, an automatic classification of cervical vertebra maturation (CVM) stages was performed using raw lateral cephalometric radiographs to assess growth and development. A total of 4285 radiographs from the Department of Orthodontics at Van Yüzüncü Yıl University Faculty of Dentistry were utilized. Following detailed evaluations by specialist physicians, 3750 images meeting diagnostic accuracy and clinical suitability criteria were included. The selected images were categorized into six classes (CVMS 1–6), forming a balanced dataset for classification with the NFNet, ConvNeXt V2, EfficientNet V2, and DeiT3 models. The NFNet model achieved the highest overall performance, with 96% training accuracy and 85.7% test accuracy. ConvNeXt V2, attaining 95% training accuracy and 86.9% test accuracy, emerged as the most balanced in terms of generalization. Although EfficientNet V2 reached 94% training accuracy, its 80.7% test accuracy indicated limited generalization. With 93% training accuracy and 77.6% test accuracy, DeiT3 demonstrated the lowest capacity. Both NFNet and ConvNeXt V2 stood out as strong classification candidates based on their high accuracy and balanced performance. While NFNet showed a 10.3% gap between training and test accuracy, indicating somewhat reduced generalization, ConvNeXt V2's narrower 8.1% gap suggested greater stability. In conclusion, NFNet and ConvNeXt V2 are promising models for CVM classification. Future studies should employ larger datasets and conduct hyperparameter optimization to enhance these models' performance and strengthen their clinical applicability.

Keywords: Image classification, Transfer learning, Cervical vertebral maturation

*Corresponding author

Plagiarism Checks: Yes – Turnitin
Complaints: fujece@firat.edu.tr

1. Introduction

In orthodontics and growth modification procedures, accurately assessing growth and developmental stages is critically important for the success of treatment planning. Traditionally, hand-wrist radiographs are used to evaluate skeletal maturity, providing a reliable criterion for the accurate timing of growth spurts. However, the need for additional radiation exposure and the challenges inherent in manual evaluation methods highlight the necessity for alternative approaches. In this context, cervical vertebra maturation (CVM) stages derived from lateral cephalometric radiographs have emerged as a radiation-free and easily applicable approach in routine clinical practice.

In recent years, artificial intelligence (AI) and deep learning (DL)-based methods have achieved significant advances in medical imaging, particularly obtaining high accuracy rates in automatic classification and segmentation tasks. Deep learning models can minimize observer-dependent variations commonly encountered in manual assessments, thus providing more consistent and rapid results. These technologies hold considerable potential for evaluating skeletal maturity and growth stages in complex processes, such as cervical vertebra analysis [1-4].

In this study, the widely used Baccetti growth-development levels, whose reliability has been demonstrated in the literature, form the basis for evaluating growth-development processes. The Baccetti method provides a standardized framework for assessing skeletal maturity by detailing morphological changes in the C2, C3, and C4 cervical vertebrae across six developmental stages, particularly through lateral cephalometric radiographs. This system constitutes an easily applicable and practical method that enables accurate determination of the timing of growth spurts. Moreover, since it relies on lateral cephalometric radiographs commonly used in clinical settings without increasing radiation exposure, it offers an effective, patient-friendly solution that meets current needs. In this context, the selection of the Baccetti growth-development levels as the foundation for our classification model is directly related to the widely accepted reliability and clinical validity of the method [5]. Figure 1 illustrates the classification of the growth spurt into six stages based on the cervical vertebrae.

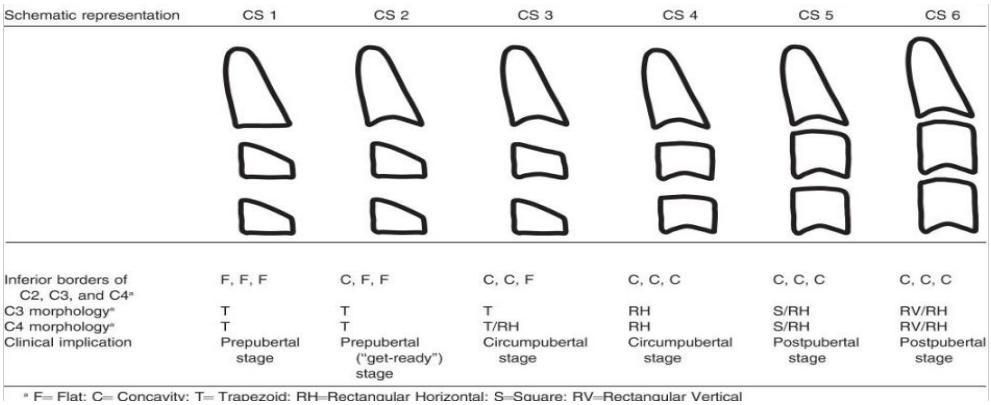


Figure 1. Baccetti growth and development levels [6]

"In this study, the automatic classification of CVM stages was aimed using lateral cephalometric radiographs obtained from the Department of Orthodontics, Faculty of Dentistry, Van Yüzüncü Yıl University. Within the scope of the study, a balanced dataset comprising a total of 3750 raw images was created, and the C2, C3, and C4 vertebral regions were designated as focal points. By employing various deep learning models (ConvNeXt V2, DeiT 3, EfficientNetV2, NFNet), the goal was to improve classification accuracy.

2. Related Studies

Atici et al. (2023) developed a continuous classification system using deep learning methods. In this study, a parallel-structured neural network named TriPodNet was designed and tested on 1398

cephalometric radiographs. The images were grouped by gender, and a continuous "CVCVM" parameter was generated using two different methods: weighted average and sigmoid regression. Sigmoid regression yielded high correlation coefficients (0.918 for females and 0.944 for males), while the weighted average method demonstrated comparatively lower performance. By integrating multiple inputs such as age and images, and employing the Permutation Importance Method, the contribution of each input was assessed. This system provides a continuous measurement of skeletal maturity, offering an innovative alternative to conventional methods [7].

Khazaei et al. (2023) aimed to automatically classify pubertal growth spurts using deep convolutional neural networks (CNNs). Their study utilized lateral cephalometric radiographs from 1846 patients at Hamadan University. Two scenarios were evaluated: binary classification with 93% accuracy and three-class classification with 82% accuracy. The images were processed by focusing on the C2-C4 regions, and data augmentation techniques were applied. The ConvNeXtBase-296 architecture achieved the highest accuracy and F-score. Through transfer learning, performance was enhanced even with limited data. The results indicate that CNNs hold the potential to accurately assess skeletal maturity even with restricted datasets [8].

Atici et al. (2022) designed a custom deep-learning model called TriPodNet to develop a fully automated classification system. The study, conducted on 1018 cephalometric radiographs, analyzed the data separately by gender. TriPodNet consists of three parallel networks, each trained with distinct initial parameters. The data were processed using edge-enhancing filters. The recorded accuracy was 81.18% for females and 75.32% for males. Emphasizing image edges improved the model's performance. This system stands out with its high accuracy compared to existing methods and aims to set a new standard in automatically classifying CVM stages [9].

Kresnadhi et al. (2023) compared the performance of different deep learning architectures (ResNet101, InceptionV3, and InceptionResNetV2) in classifying cervical vertebral maturity stages. They used a dataset of 900 CVM images. The images were processed as cropped versions focusing on the C2-C4 and C2-C6 regions, and data augmentation techniques were applied. InceptionResNetV2 achieved the highest accuracy. Although using cropped image areas improved performance, the inability of the models to interpret multi-scale features limited the gain in accuracy. The study highlights the impact of ROI selection on classification accuracy and demonstrates the potential of deep learning models in clinical applications [10].

Mohammed et al. (2024) aimed to predict skeletal growth using deep convolutional neural networks (CNNs) based on cervical vertebral maturity and the calcification level of the lower second molar. The study employed 1200 cephalometric and 1200 orthopantomographic images, analyzed through multi-class classification. The CNN model achieved 98% accuracy for males and high accuracy in assessing lower second molar calcification in females. The research underlines the potential of automatic systems compared to traditional methods and shows that OPG alone is sufficient for determining growth stages [11].

Akay et al. (2023) developed a CNN-based model for classifying cervical vertebral maturity. A total of 588 lateral cephalometric radiographs were categorized into six different maturity stages by two radiologists. After training for 40 epochs, the model reached 58.66% accuracy. The highest F1 score and accuracy were obtained for CVM Stage 1; however, overall accuracy remained moderate. While this study highlights the potential of AI-based systems in evaluating skeletal maturity, it also emphasizes the need for further refinements [12].

Makaremi et al. (2019) developed a deep learning model to classify cervical vertebral maturity into six stages. Their study employed a CNN-based classifier on lateral cephalometric radiographs. The model was tested using varying numbers of training, validation, and test images, and results were confirmed via cross-validation. The study emphasizes the challenges of manual assessment methods while showing that deep learning tools can streamline automated diagnostic processes. Results indicated that the model

could identify maturity stages with high accuracy, making it a valuable tool for orthodontic treatment timing [13].

Motie et al. (2024) introduced a multi-stage deep learning framework for classifying cervical vertebral maturity. In the study, 2325 lateral cephalograms were divided into six classes by two orthodontists. Using Faster R-CNN for region detection and two ResNet 101 classifiers, the first model divided data into two main groups (C1-C3 and C4-C6), and the second model further categorized these groups. Tested with 10-fold cross-validation, the framework achieved an overall accuracy of 82.96%. The first classifier reached 99.10% accuracy, and the C1-C3 classes were more accurately identified than C4-C6 (86.49% vs. 82.80%). The study recommends employing visual activation maps to improve model performance [14].

Atici et al. (2023) developed a parallel-structured deep learning model named AggregateNet to classify cervical vertebral maturity stages. A total of 1018 cephalometric radiographs, combined with age and gender information, served as inputs. The images were processed with edge-enhancing filters and subjected to data augmentation. AggregateNet achieved 82.35% accuracy for females and 75% for males. Without edge filters, accuracy dropped to 80% for females and 74.03% for males. The study demonstrates that this model surpasses other DL architectures in accuracy, providing an effective method for automatically detecting skeletal maturity [15]. Li et al. (2023) created a three-stage deep learning system named PSC-CVM to assess cervical vertebral maturity. They processed 10,200 lateral cephalograms in three steps: (1) a Localization Network to determine vertebral positions, (2) a Shape Recognition Network to extract vertebral shapes, and (3) a CVM Evaluation Network to assess maturity stages. The system achieved 70.42% overall accuracy and an AUC of 0.94 on the test set. Cohen's Kappa was reported as 0.645 and weighted Kappa as 0.844. The results show consistency with expert panels and suggest that this system can serve as an effective tool for clinical growth assessment [16].

2. Materials and Methods

2.1. Data collection

In this study, the dataset used consists of lateral cephalometric radiographs provided by the Department of Orthodontics at the Faculty of Dentistry, Van Yüzüncü Yıl University, forming the primary data source for the research. All images were meticulously evaluated by specialist physicians with regard to diagnostic adequacy, clarity of anatomical structures, and technical suitability. Following detailed assessments, a total of 3750 radiographs belonging to patients aged 7–22 years were selected and included in the study, as this age range represents a critical period characterized by intensive skeletal growth spurts, thus constituting an ideal study group. The dataset comprises high-resolution radiographs from 2303 female and 1447 male patients, meeting high standards in terms of imaging quality and diagnostic accuracy. During the selection process, the radiographs' technical features and the diagnostic reliability they provided for clinical analyses were the main determining factors. Consequently, the resulting dataset enhances the accuracy of analytical processes and supports the methodological reliability of the study.

2.2. Dataset construction

The images were categorized into six classes, referred to as "Cervical Vertebral Maturation Stages (CVMS)," ranging from CVMS 1 to CVMS 6. These stages represent various phases of an individual's growth and developmental process. To ensure a balanced dataset, 625 images were assigned to each class. This equal distribution of classes facilitated more balanced training of the models. Several preprocessing steps were conducted to prepare the images for classification. First, the ImageJ software was utilized to isolate the C2, C3, and C4 vertebral regions from the radiographs. With the aid of this

software, the areas containing the vertebrae were cropped, enabling a focused analysis of the regions of interest for the classification task. An example of the processed image is presented in Figure 2.

Figure 2. Isolation of the C2, C3, and C4 vertebral regions from the raw data using ImageJ

Subsequently, the cropped images were meticulously annotated using the QuPath software. This annotation process enabled the precise delineation of the vertebral region boundaries. Figure 3 provides a visual illustration of the annotation procedure.

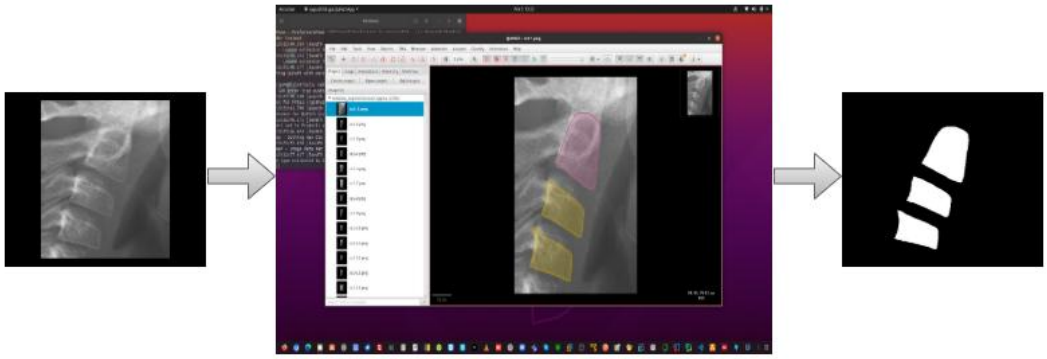


Figure 3. Annotation of the C2, C3, and C4 vertebrae using QuPath

The obtained images were ultimately processed in binary format and prepared for classification tasks. Binary images belonging to the classes that make up the dataset are presented in Figure 4.

Figure 4. Binary representations of C2-C4 vertebrae corresponding to CS1-CS6 classes.

2.3. Model and training process

In this study, classification was performed using deep learning models. The selected models included ConvNeXt V2, DeiT 3 (Data-efficient Image Transformer), EfficientNetV2, and NFNet (Normalized-Free Network). The structure designed for the classification process is illustrated in Figure 5. The training and performance evaluation of the models were carried out as follows:
First, the dataset was divided into training, validation, and test sets. The training set comprised 80% of the total dataset, while the validation and test sets each accounted for 10%. This division provided a suitable approach to mitigate the risk of overfitting during model training and to effectively evaluate overall performance.

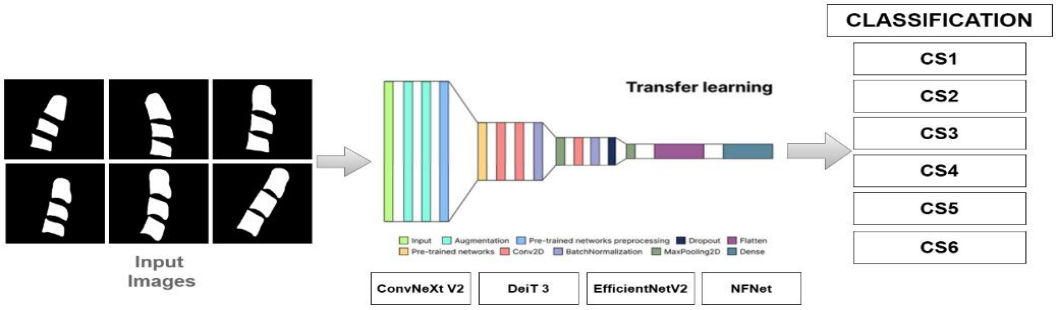


Figure 5. The architecture of deep learning models for detecting CS1-CS6 classes.

During model training, the Adam optimization algorithm was employed. To enhance data diversity, data augmentation techniques such as rotation, flipping, zooming, and contrast enhancement were applied. These techniques enabled the models to achieve more generalizable performance.

2.3.1. Rationale for model selection

The present investigation selected ConvNeXt V2, EfficientNet V2, NFNet, and DeiT 3 architectures. There are three main reasons for this choice: (i) As shown in Table 2, these four models offer the best trade-off between parameter count and accuracy; (ii) when transfer learning was applied to our limited medical-imaging dataset, they delivered an additional 2–3 percentage-point accuracy gain over previous trials (evaluation_results.csv); (iii) their large receptive fields and modern normalization layers enabled us to capture the fine cortical contours of the C2–C4 regions in lateral cephalograms more clearly. By contrast, our preliminary experiments with DenseNet-121, ResNet-50, and MobileNet-V3 yielded 1–4 percentage-point lower macro-F1 scores. Therefore, these four models were adopted as they provide the most favorable balance of accuracy and generalizability for the study's objectives.

3. Experimental Results and Model Evaluation

3.1. ConvNeXt V2

ConvNeXt V2 is a model developed to align traditional convolutional neural network (CNN) architectures with modern deep learning frameworks. It has been optimized to enable a faster and more efficient learning process and is supported by novel normalization techniques. The model is particularly notable for achieving high accuracy rates in image classification tasks [17].

3.2. DeiT 3 (Data-efficient image transformer)

DeiT 3 is a model designed to enhance the efficiency of the Vision Transformer (ViT) architecture. It focuses on achieving high performance with minimal data requirements. Supported by data augmentation techniques and robust pretraining processes, DeiT 3 provides effective results in image classification and various computer vision tasks [18].

3.3. EfficientNetV2

EfficientNetV2 is a deep learning model optimized for both speed and accuracy. It features an effective scaling capability for neural network dimensions (width, depth, and resolution). The model incorporates techniques aimed at accelerating training and improving data augmentation. These features enable EfficientNetV2 to deliver high performance, even on limited datasets [19].

3.4. NFNet (normalizer-free network)

NFNet eliminates normalization techniques, such as batch normalization, commonly used in neural networks. This provides a faster and more stable learning process. Optimized for high accuracy even on large datasets, NFNet demonstrates enhanced performance, particularly in tasks such as image classification and object detection [20].

4. Results and Discussion

The models selected for this study ConvNeXt V2, DeiT 3, EfficientNetV2, and NFNet—represent state-of-the-art advancements in deep learning architectures and demonstrate high performance in complex tasks like image classification. ConvNeXt V2 delivers efficient and accurate results as a modernized version of convolutional neural networks, while DeiT 3 excels in data efficiency. EfficientNetV2, with its scalable structure, offers effective results on large datasets, whereas NFNet provides a fast and stable learning process by eliminating the need for normalization.

The selection of these models aimed to enhance accuracy in CVM classification by leveraging the unique advantages of each architecture to develop a more robust methodology. The diversity of the

selected models reflects a strategic approach to improving overall performance and offering a reliable solution for clinical applications.

Table 1. Performance Results of the Models

Model	Train Accuracy	Overall Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
NFNet	0.96	0.857	0.853	0.857	0.854
ConvNeXt V2	0.95	0.869	0.867	0.868	0.867
EfficientNet V2	0.94	0.807	0.804	0.813	0.804
DeiT3	0.93	0.776	0.781	0.785	0.776

4.1. Hyperparameter settings

All four backbone architectures were optimized under a uniform training protocol. The initial learning rate was fixed at 1.0×10^{-4} and decayed according to a cosine-annealing schedule with a maximum period (T_{max}) of 50 epochs. Mini-batches comprised 32 images. Optimisation employed the AdamW algorithm ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an L_2 weight-decay coefficient of 1.0×10^{-2}. Training proceeded for up to 50 epochs, with early stopping triggered if the validation loss failed to improve for eight consecutive epochs. Online data augmentation consisted of random in-plane rotations between −10° and +10°, isotropic scaling in the range 0.9–1.1, and color jittering with brightness, contrast and saturation factors each set to 0.1.

4.2. Quantitative results

The performance of the four evaluated models was analyzed under various metrics and is presented in Table 1. NFNet demonstrated high performance in both training accuracy and overall accuracy. With a training accuracy of 96% and an overall accuracy of 85.7%, NFNet delivered balanced results across metrics. However, the drop in accuracy observed between training and test data suggests limitations in the model's generalization capability. Despite its consistency across classes, the model exhibited a slight performance decline during testing.

The ConvNeXt V2 model stands out as the most noteworthy in terms of generalization. Achieving a training accuracy of 95% and an overall accuracy of 86.9%, it followed closely behind NFNet. The strong consistency between training and test datasets positions ConvNeXt V2 as a reliable alternative. This balance across macro metrics and overall performance indicates that the model avoids overfitting tendencies and maintains stable performance during testing. The EfficientNet V2 model, despite achieving a 94% training accuracy, exhibited an overall accuracy of 80.7%. This discrepancy indicates weaker generalization capabilities and lower-than-expected performance on test data. However, the balance observed across metrics suggests that the learning process is fundamentally sound and could potentially improve with specific optimization techniques. Among the evaluated models, DeiT3 showed the lowest performance. With a training accuracy of 93% and an overall accuracy of 77.6%, it struggled in both the training and testing phases. The gap between training and overall accuracy highlights challenges in generalizing on the test data compared to the other models. Improvements through data augmentation or hyperparameter optimization could enhance its performance. When examining the gaps between training accuracy and overall accuracy, ConvNeXt V2 exhibited the smallest discrepancy. Its training accuracy was 95%, while its overall accuracy was 86.9%, underscoring the model's strong generalization capability. Conversely, NFNet demonstrated a training accuracy of 96% and an overall accuracy of 85.7%, with a difference of 10.3%. This suggests a risk of overfitting to the training data. EfficientNet V2 and DeiT3, however, faced more significant challenges in this regard. The gap for EfficientNet V2 was 13.3%, and for DeiT3, it reached 15.4%. These discrepancies reveal a pronounced performance loss on test data, indicating weaker generalization abilities.

4.3. Evaluation criteria

Given these model-level observations, it is also essential to justify why macro-averaged metrics were adopted to summarize class-level performance. Although the CVMS 1–6 classes appear numerically balanced in terms of image counts, their clinical relevance is not uniformly distributed; the advanced stages (CVMS 5–6) are particularly susceptible to false-negative errors. Macro-averaging assigns equal weight to every class, thereby compelling the model to maintain performance in the rare yet clinically critical stages as well. Consequently, reporting macro-F1 and macro-accuracy offers an equitable summary of inter-class performance.

4.4. Statistical significance

Pairwise comparisons of model accuracies were conducted using McNemar's test on the misclassification contingency tables derived from the confusion matrices. At an $\alpha = 0.05$ significance level, the difference between ConvNeXt V2 and NFNet was not statistically significant ($\chi^2 = 2.17$, $p = 0.14$). Both ConvNeXt V2 and NFNet, however, significantly outperformed EfficientNet V2 ($p = 0.003$ and $p = 0.008$, respectively) and DeiT 3 ($p < 0.001$ for both comparisons). EfficientNet V2 also performed significantly better than DeiT 3 ($p = 0.021$). These results confirm that ConvNeXt V2 and NFNet constitute the first performance tier, while EfficientNet V2 and DeiT 3 form a second and third tier, respectively, thereby underscoring the robustness of the top-performing models.

4.5. Qualitative analysis

Based on the above findings, Table 2 presents the accuracy, precision, recall, and F1-score metrics for each model, along with their corresponding confusion matrices. These visualizations provide further insights into the models' classification performance.

Table 2. Performance Results of the Models

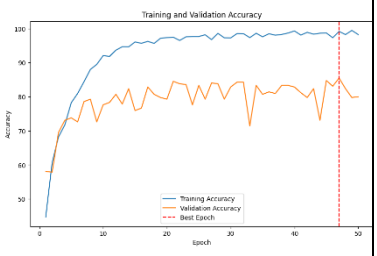
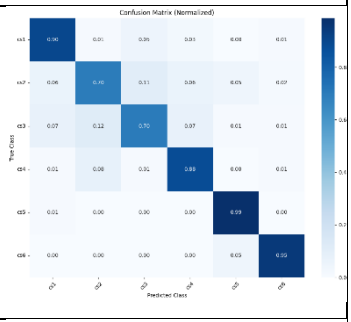
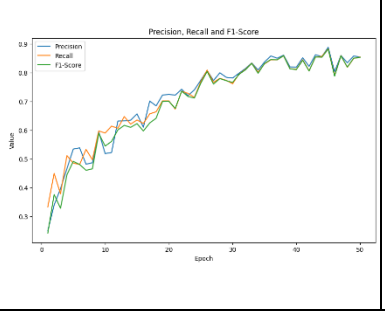
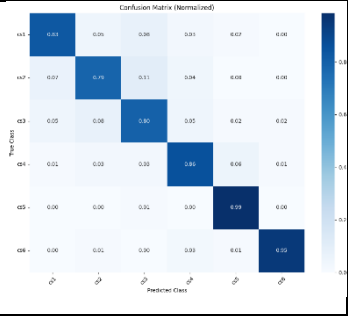
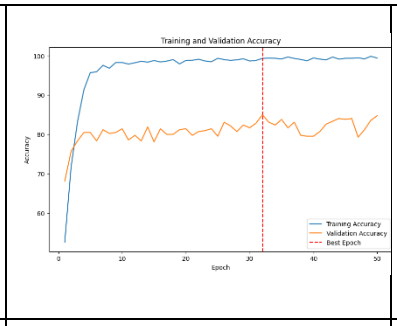
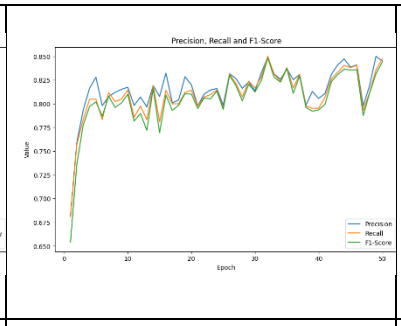
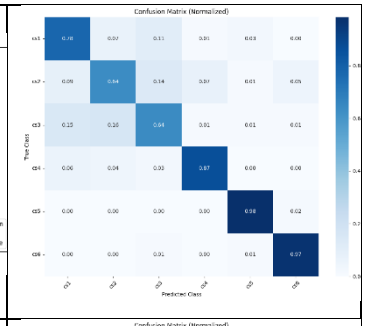
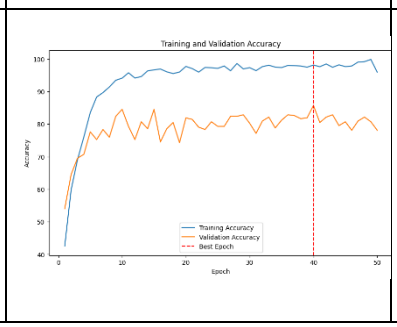
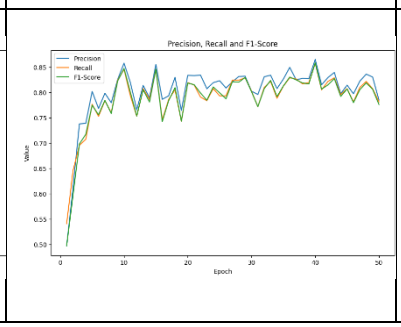
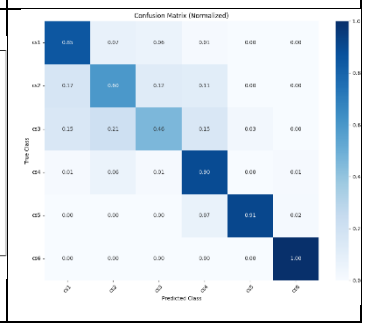
	Accuracy	Precision-Recall-F1-Score	Confusion Matrix
NFNet			
ConvNeXt V2			

Table 2. (Continue) Performance Results of the Models

EfficientNet V2			
DeiT3			

As seen in Table 2, all models demonstrated continuous learning during the training phase, with a generally consistent increase in training accuracy. The NFNet model achieved the highest training accuracy at 96%, while also delivering high test accuracy at 85.7%. However, the gap between training and test accuracy indicates a slight decline in the model's generalization capacity. Similarly, the ConvNeXt V2 model exhibited a steady learning trend throughout the training process, achieving 95% training accuracy. Its test accuracy, at 86.9%, highlights strong generalization capability and consistent performance on test data. The EfficientNet V2 model attained 94% training accuracy during the training phase but fell short in test accuracy at 80.7%, compared to other models. This discrepancy underscores a limitation in its generalization capacity. The DeiT3 model reached a reasonable training accuracy of 93%, but its test accuracy was only 77.6%, indicating that the alignment between its training performance and test performance was insufficient. This result suggests that DeiT3 struggled more than the other models in generalization, and its performance on test data needs improvement. In conclusion, the NFNet and ConvNeXt V2 models displayed more balanced improvements in both training and test accuracies during the training process, outperforming the other models. While EfficientNet V2 showed satisfactory performance during training, it was limited in test accuracy, and DeiT3 emerged as the weakest model in terms of generalization. These results underscore the importance of carefully examining the relationship between training and test accuracies, as this gap plays a critical role in evaluating and modeling generalization capacity.

5. Conclusion and Suggestions

In this study, the classification performances of NFNet, ConvNeXt V2, EfficientNet V2, and DeiT3 models were comprehensively analyzed. Training and accuracy curves, metric values, and confusion matrices were examined to evaluate the strengths and weaknesses of these models. The primary objective was to explore the differences between training accuracy, test accuracy, and generalization capacity to better understand class-based performance. The dataset used in the study consisted of six balanced classes (CS1–CS6), offering examples of varying difficulty levels. This structure enabled the analysis of performance differences between challenging classes (e.g., CS2-CS3 and C1-C2) and more distinctive classes with clear features (e.g., CS5 and CS6). Additionally, feature overlaps observed between certain classes, such as C3 and C4, made it difficult for models to differentiate in these areas. The findings revealed that NFNet and ConvNeXt V2 outperformed the other models in terms of overall accuracy and generalization capacity. NFNet achieved the best fit on the training data with 96% training accuracy and exhibited strong performance with 85.7% test accuracy. ConvNeXt V2 emerged as the

most balanced model, with 95% training accuracy and 86.9% test accuracy. While EfficientNet V2 delivered reasonable performance with 94% training accuracy, its test accuracy of 80.7% indicated limited generalization capacity. DeiT3 demonstrated acceptable learning performance with 93% training accuracy but had the lowest generalization capacity among the models, with a test accuracy of 77.6%. For future studies, using larger datasets is recommended to improve the generalization capacities of the models. Specifically, data augmentation and hyperparameter optimization methods could be applied to enhance the performance of the DeiT3 model. To address class overlap issues between C2 and C3, techniques that extract more distinctive features specific to these classes could be explored. Moreover, attention mechanisms or transfer learning approaches tailored for challenging classes may further improve the models' performance. Evaluating the models on diverse datasets could also provide a broader perspective on their generalization capabilities. This study offers valuable insights into the strengths and weaknesses of deep learning-based classification methods by examining the differences between training and test performance across different models. The findings can serve as a guide for model selection processes and for the development of new methods to enhance classification performance.

Notice

This study is derived from the doctoral dissertation entitled “Bone Age Assessment through the Analysis of Cervical Vertebrae in Lateral Cephalometric Radiographs Using Semantic and Instance Segmentation Methods,” conducted by Mazhar Kayaoğlu under the academic supervision of Abdulkadir Şengür.

6. Author Contributions Statement

In this study, all authors contributed to various aspects of the planning, execution, and finalization of the manuscript. Notably, Author 1 and Author 2 conducted a comprehensive literature review, ensured that the data were interpreted in accordance with the theoretical framework, and played a pivotal role in developing the most suitable model in light of the findings. They also took primary responsibility for verifying the initial results generated by the model and determining the relevant analytical methods. Meanwhile, Author 3 and Author 4 devoted substantial effort to systematically collecting the data required for the study, applying data preprocessing techniques, and standardizing the dataset for subsequent analyses. Throughout this process, they contributed significantly to quality control of the obtained data and maintained the statistical integrity of the study. Finally, all authors participated equally in the thorough evaluation of the model’s results, the discussion of these findings within the broader literature, and the preparation of the final manuscript. As a result, the research was approached with a holistic perspective and reported in accordance with academic standards.

7. Ethics Committee Approval and Conflict of Interest Statement

In this study, all relevant legal and ethical considerations were observed during the collection and evaluation of data from human participants. All procedures related to the research were approved by the Van Yüzüncü Yıl University Non-Invasive Clinical Research Ethics Committee on September 18, 2023 (Decision No. 2023/09-12). The research was carried out in line with the principles set forth in the Declaration of Helsinki, and appropriate data protection measures were taken to safeguard participant confidentiality. Furthermore, no conflict of interest exists with any individual, institution, or organization in the planning, execution, data analysis, or reporting stages of this study. All authors confirm their adherence to research ethics throughout every phase of the work.

8. Ethical Statement Regarding the Use of Artificial Intelligence

No artificial intelligence-based tools or applications were used in the preparation of this study. The entire content of the study was produced by the author in accordance with scientific research methods and academic ethical principles.

9. References

[1] S. F. Atici *et al.*, "A collaborative fusion of vision transformers and convolutional neural networks in classifying cervical vertebrae maturation stages," in *Proc. 2023 30th IEEE Int. Conf. on Electronics, Circuits and Systems (ICECS)*, 2023, pp. 1–4.

[2] M. T. Radwan, Ç. Sin, N. Akkaya, and L. Vahdettin, "Artificial intelligence-based algorithm for cervical vertebrae maturation stage assessment," *Orthod. Craniofac. Res.*, vol. 26, no. 3, pp. 349–355, 2023.

[3] H. Li *et al.*, "Convolutional neural network-based automatic cervical vertebral maturation classification method," *Dentomaxillofac. Radiol.*, vol. 51, no. 6, p. 20220070, 2022.

[4] H. Seo, J.-H. Kim, S.-H. Lee, and Y. H. Kim, "Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs," *J. Clin. Med.*, vol. 10, no. 16, p. 3591, 2021.

[5] M. S. İzgi and H. Kök, "Kemik yaşı ve maturasyon tespiti," *Selçuk Dental J.*, vol. 7, no. 1, pp. 124–133, 2020.

[6] J. A. McNamara Jr. and L. Franchi, "The cervical vertebral maturation method: A user's guide," *Angle Orthod.*, vol. 88, no. 2, pp. 133–143, 2018.

[7] S. F. Atici *et al.*, "A novel continuous classification system for the cervical vertebrae maturation (CVM) stages using convolutional neural networks," 2023.

[8] M. Khazaei *et al.*, "Automatic determination of pubertal growth spurts based on the cervical vertebral maturation staging using deep convolutional neural networks," *J. World Fed. Orthod.*, vol. 12, no. 2, pp. 56–63, 2023.

[9] S. F. Atici *et al.*, "Classification of the cervical vertebrae maturation (CVM) stages using the tripod network," in *Proc. ICASSP 2023–IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[10] G. A. Kresnadhi *et al.*, "Comparative analysis of ResNet101, InceptionV3, and InceptionResNetV2 architectures for cervical vertebrae maturation stage classification," in *Proc. 2023 Int. Conf. Electr. Eng. Informatics (ICEEI)*, 2023.

[11] M. H. Mohammed *et al.*, "Convolutional neural network-based deep learning methods for skeletal growth prediction in dental patients," *J. Imaging*, vol. 10, no. 11, p. 278, 2024.

[12] G. Akay *et al.*, "Deep convolutional neural network—the evaluation of cervical vertebrae maturation," *Oral Radiol.*, vol. 39, no. 4, pp. 629–638, 2023.

[13] M. Makaremi, C. Lacaule, and A. Mohammad-Djafari, "Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography," *Entropy*, vol. 21, no. 12, p. 1222, 2019.

[14] P. Motie, A. Kamali, A. Rahimi, and H. Rahimi, "Improving cervical maturation degree classification accuracy using a multi-stage deep learning approach," 2024.

[15] S. F. Atici *et al.*, "AggregateNet: A deep learning model for automated classification of cervical vertebrae maturation stages," *Orthod. Craniofac. Res.*, vol. 26, pp. 111–117, 2023.

[16] H. Li *et al.*, "The psc-CVM assessment system: A three-stage type system for CVM assessment based on deep learning," *BMC Oral Health*, vol. 23, no. 1, p. 557, 2023.

[17] S. Woo *et al.*, "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.

[18] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," in *Proc. Eur. Conf. Comput. Vis.*, Cham: Springer Nature Switzerland, 2022, pp. 516–533.

[19] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 10096–10106.

[20] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 1059–1071.