Joint Diseases and Related Surgery

Jt Dis Relat Surg
2023;34(3):598-604

ORIGINAL ARTICLE

Automated fracture detection in the ulna and radius using deep learning on upper extremity radiographs

Çağatay Berke Erdaş

Department of Computer Engineering, Faculty of Engineering, Başkent University, Ankara, Türkiye

ABSTRACT

Objectives: This study aimed to detect single or multiple fractures in the ulna or radius using deep learning techniques fed on upper-extremity radiographs.

Materials and methods: The data set used in the retrospective study consisted of different types of upper extremity radiographs obtained from an open-source dataset, with 4,480 images with fractures and 4,383 images without fractures. All fractures involved the ulna or radius. The proposed method comprises two distinct stages. The initial phase, referred to as preprocessing, involved the removal of radiographic backgrounds, followed by the elimination of nonbone tissue. In the second phase, images consisting only of bone tissue were processed using deep learning models, such as RegNetX006, EfficientNet B0, and InceptionResNetV2. Thus, whether one or more fractures of the ulna or the radius are present was determined. To measure the performance of the proposed method, raw images, images generated by background deletion, and bone tissue removal were classified separately using RegNetX006, EfficientNet B0, and InceptionResNetV2 models. Performance was assessed by accuracy, F1 score, Matthew's correlation coefficient, receiver operating characteristic area under the curve, sensitivity, specificity, and precision using 10-fold cross-validation, which is a widely accepted technique in statistical analysis.

Results: The best classification performance was obtained with the proposed preprocessing and RegNetX006 architecture. The values obtained for various metrics were as follows: accuracy (0.9921), F1 score (0.9918), Matthew's correlation coefficient (0.9842), area under the curve (0.9918), sensitivity (0.9974), specificity (0.9863), and precision (0.9923).

Conclusion: The proposed preprocessing method is able to detect fractures of the ulna and radius by artificial intelligence.

Keywords: Deep learning, fracture detection, ulna and radius.

Fracture detection is of utmost importance in medical diagnosis and treatment planning, particularly when it comes to upper extremity injuries. In the past, this task used to heavily rely on manual examination of radiographs by medical experts, which was a time-consuming and error-prone process. However, thanks to exciting advancements in deep learning techniques and the abundance of extensive medical imaging, datasets automated fracture detection has witnessed a transformative revolution.[1]

Automated fracture detection systems have the potential to significantly reduce the workload of radiologists and shorten the time required for diagnosis. Instead of manually reviewing each radiograph, physicians can rely on artificial intelligence-powered algorithms to assist them with the initial examination. These systems can quickly analyze images and highlight suspicious regions, allowing radiologists to focus their attention on areas of interest and make more informed decisions. Early detection and treatment not only help preserve joints but also help patients maintain their quality of life and mobility in the postoperative period. Although artificial intelligence-assisted automated fracture detection systems can work with data obtained from various imaging modalities, radiography, which is the fastest, easiest, and cheapest method, is the most suitable.[2]

Received: July 12, 2023
Accepted: July 30, 2023
Published online: August 22, 2023

Correspondence: Çağatay Berke Erdaş. Başkent Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06790 Etimesgut, Ankara, Türkiye.

E-mail: berdas@baskent.edu.tr

Doi: 10.52312/jdrs.2023.1312

Citation: Erdaş ÇB. Automated fracture detection in the ulna and radius using deep learning on upper extremity radiographs. Jt Dis Relat Surg 2023;34(3):598-604. doi: 10.52312/jdrs.2023.1312

Vishnu et al.[3] used radiography of the humerus, ulna, femur, tibia, and fibula to diagnose the fracture type. The photographs were preprocessed using the Canny edge detection and Harris corner detector techniques. The bag-of-words model, which is extensively used in text categorization, was utilized for feature extraction. The classifier in the study was support vector machine (SVM), and the system achieved 78% performance based on 10-fold cross-validation findings.

Bayram and Çakıroğlu[4] focused on the classification of diaphyseal femur fractures using radiography. The classification process aimed to identify nine different types of fractures within the dataset, which included a total of 196 femoral radiographs. Based on performance evaluations using the 10-fold cross-validation technique, the SVM classifier was shown to outperform other classifiers, achieving a score of 89.87% on the overall accuracy criterion. Their results highlight the SVM's performance in classifying femoral fractures on radiographs.

Beyaz et al.[5] used a convolutional neural network (CNN) to detect fractures in a total of 234 frontal pelvic radiographs from 65 subjects in a dataset of radiographs, and a genetic algorithm to improve classification performance. They proved the effectiveness of the techniques they proposed with the values of 82.5% without genetic algorithm and 83.6% with confidence interval for the F1 score.

Similarly, Şahin[6] classified femoral neck fractures using a dataset consisting of a total of 193 radiographs. They used Hough and Harris methods as well as Canny and Sobel algorithms to extract more valuable features for classification. These features fed 12 different machine learning classifiers with linear discriminant analysis, which gave the highest success rate of 88.67%. They preferred a grid search approach to tune the classifier hyperparameters used in the study and obtained the final results using 10-fold cross-validation.

Rashid et al.[7] attempted to detect wrist fractures from radiographs. They proposed a new deep learning model combining extended CNN and long short-term memory. Image preprocessing and data augmentation techniques were used to overcome the difficulties that the data set with a relatively small number of samples may cause in classification. A 28-layer CNN was used for deep feature extraction, followed by a long short-term memory network for fracture identification. In the study, the highest accuracy obtained was 88.24%.

Directional information in radiographs, which contains metadata for clinicians, creates noise for artificial intelligence. A similar situation applies to nonbone tissues. Although nonbone tissues are not important for clinicians in decision-making, they are not considered as any noise. In terms of artificial intelligence, nonbone tissues can be considered as noise, as they contain insignificant additional information and directly affect the evaluation. Eliminating these metadata and nonbone tissues can improve classification performance.

These studies demonstrate the application of advanced techniques, such as image processing, genetic algorithms, and machine/deep learning models, in fracture detection. The results highlight the potential of these methods to accurately detect fractures from radiographs and provide valuable information to improve fracture diagnosis and patient care. In addition, these studies focused on the detection of a particular type of fracture in a single radiographic type rather than addressing the detection of various types of single or multiple fractures in different radiographic types and bones. This study aimed to detect single or multiple fractures in the ulna or radius using deep learning techniques fed on upper-extremity radiographs.

MATERIALS AND METHODS

Dataset

The retrospective study utilized a diverse dataset of upper extremity radiographs, which was preferred as an open-source dataset brought to the literature.[8] This dataset included lateral and anteroposterior radiographs of the forearm, lateral and posteroanterior radiographs of the wrist, and lateral and anteroposterior radiographs of the elbow. These images documented various types of fractures, including oblique, transverse, greenstick, comminuted, segmental, and spiral fractures. Overall, the dataset consisted of 8,863 images. Out of these images, 4,480 exhibited fractures located at the ulna or radius. The remaining 4,383 images depicted nonfractured cases.

General framework

The method proposed in this study consists of two stages, namely preprocessing, which includes the elimination of metadata and nonbone tissues, and artificial intelligence-based classification.

Adaptive Background Removal with Edge Attention algorithm, which is a background removal algorithm, was used to delete metadata in

radiographs. The Adaptive Background Removal with Edge Attention algorithm efficiently removes image backgrounds by adapting to background characteristics and utilizing edge information. It operates hierarchically through multiple stages and convolutional layers.[9] The proposed method of elimination of nonbone tissue includes creating negative states of radiographs, bounded sum operation, and re-negative to convert the image back to the original color space after bounded sum operation, respectively. The images that are cleared of metadata and nonbone tissue after these processes are called preprocessed images. An example illustration of the proposed preprocessing technique is given in Figure 1.

As can be seen in Figure 2, preprocessed images were divided into two parts for testing and training. While the CNN models used in this study were trained with the examples in the subdataset reserved for training, the models' classification performance for the detection of fracture or fractures in the ulna and radius bones was tested using the remaining examples.

Methods

To eliminate nonbone tissue, the image with the erased background was converted to negative and the bounded sum method was applied. The resulting negative image underwent another bounded sum operation and was then restored to its original color space. This process effectively removes nonbone tissues while preserving the bone structure. Otsu thresholding was used to determine the optimal cutoff value for each radiograph, ensuring the best results for each image.

Three architectures were utilized in this study: RegNetX006, EfficientNet B0, and InceptionResNetV2. RegNetX006's modular structure

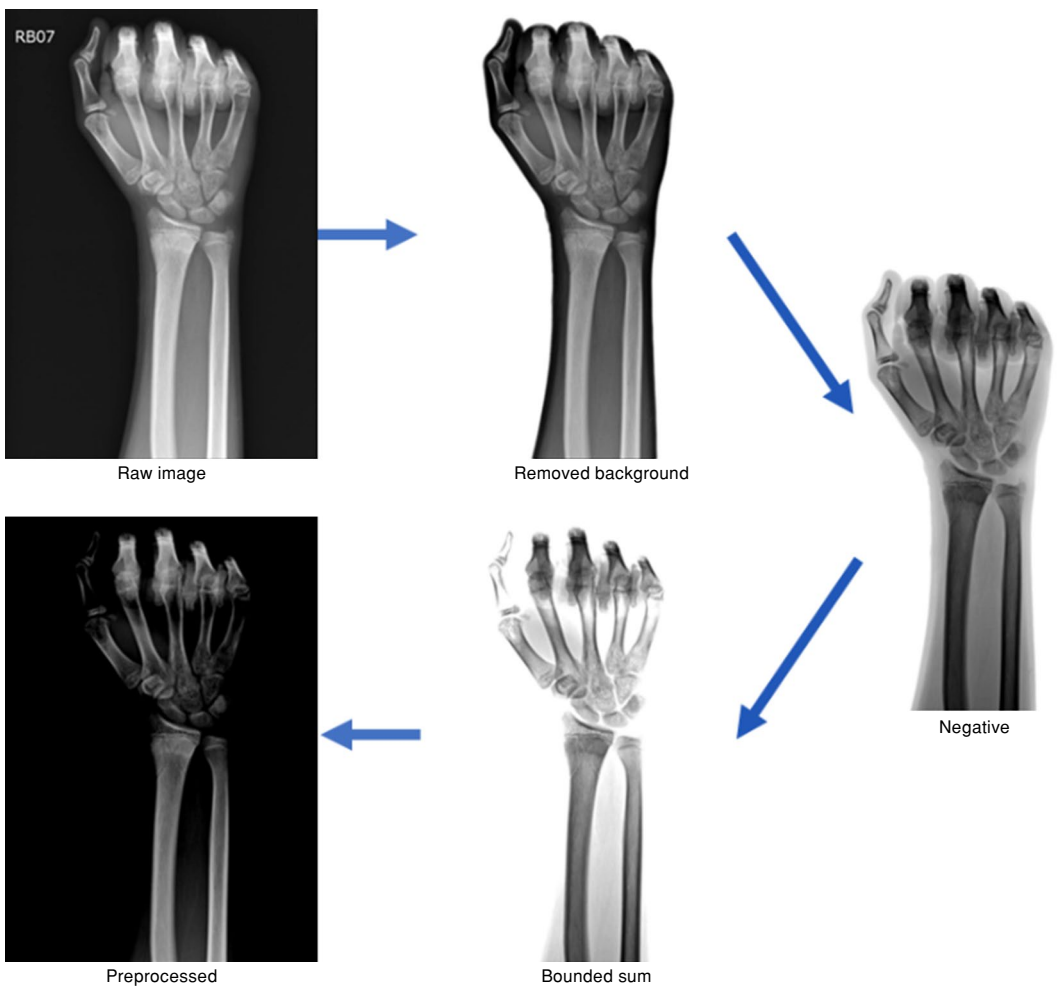

FIGURE 1. Proposed preprocessing steps.

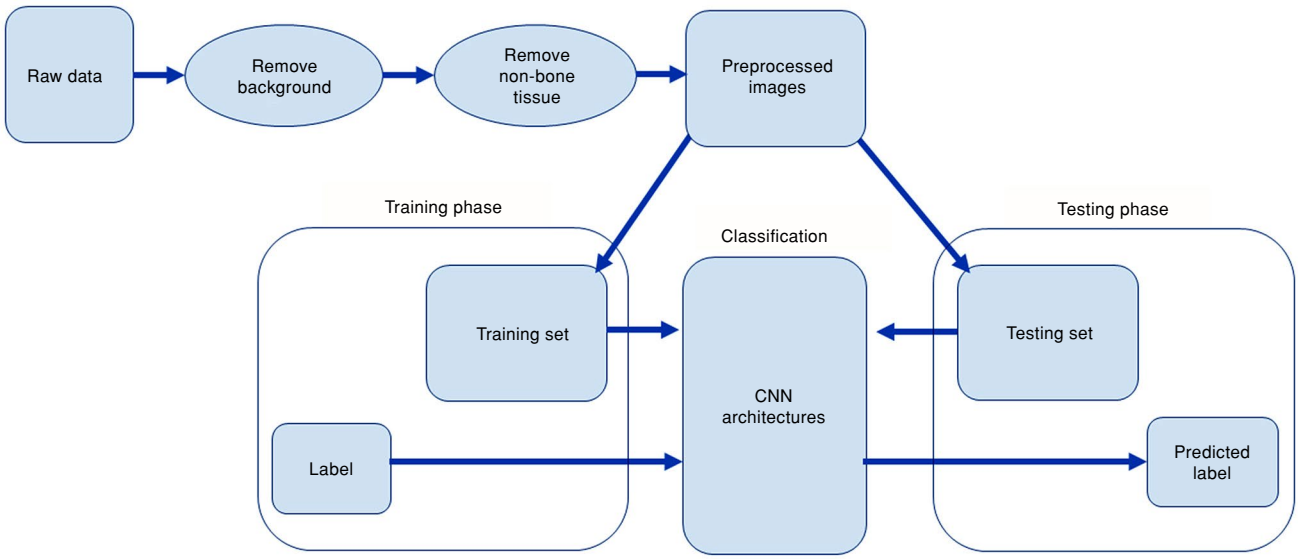

FIGURE 2. General overview of the study.
CNN: Convolutional neural network.

enhances generalization capabilities and prevents overfitting with multiple stages, blocks, and groups. Down-sampling operations accommodate varying object sizes and skip connections aid efficient training. The stages capture complex features, while blocks refine learned features using multiple convolutional layers. These transformations enhance learning and representation.[10] EfficientNet B0 is a computationally efficient model that optimizes depth, width, and resolution using a composite scaling strategy. It features inverted bottleneck blocks with depth-separable convolutions, reducing parameters and computational complexity while improving performance. This design allows for effective spatial information capture and efficient channel compression, achieving a balance between computational efficiency and accurate representation.[11] The InceptionResNetV2 is an advanced deep convolutional neural network that merges both the Inception and ResNet architectures. This powerful model incorporates the use of inception modules and residual connections, which are arranged in repeated blocks to form a deep network.[12]

RESULTS

Performance evaluation

To obtain accurate performance assessment, the models were evaluated using 10-fold cross-validation.[13] A batch size of 64 was used, along with a learning rate of 0.01 and 10 epochs.

The performance of the system was assessed using various metrics, such as accuracy, F1 score, Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC), sensitivity, specificity, and precision.[13] As can be seen in Equation 1, accuracy indicates how frequently the system accurately predicts outcomes and is determined by dividing the number of correct predictions by the total number of predictions made.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP+FP+TN+FN)} \quad (1)$$

The F1 score is a measurement of overall performance that considers both precision and recall. It provides a balanced assessment by considering false positive (FP) and false negative (FN) predictions. The F1 score, calculated as the harmonic mean of precision and recall, is formulated in Equation 2.

$$\text{F1 Score} = \frac{2TP}{(2TP+FP+FN)} \quad (2)$$

The MCC takes into consideration true positive (TP), true negative (TN), FP, and FN predictions to produce a balanced performance metric that is especially beneficial for unbalanced data sets (Equation 3).

$$\text{MMC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \quad (3)$$

The ROC curve is a visual depiction of how well the system performs at various classification thresholds. The area under the curve (AUC) is used to gauge the overall performance of the system. A higher AUC indicates that the system is better able to distinguish between positive and negative instances.

The fraction of TP cases successfully detected by the system is measured by sensitivity, also known as recall. Sensitivity, which is calculated by dividing the number of TPs by the total number of TP and FN cases, is formulated in Equation 4.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP+FP})} \quad (4)$$

Specificity measures the proportion of TN instances that were correctly identified by the system. It is calculated by dividing the TNs by the sum of the TNs and FPs (Equation 5).

$$\text{Specificity} = \frac{\text{TN}}{(\text{FP+TN})} \quad (5)$$

Precision indicates how many of the predicted positive instances are actually positive. It is calculated by dividing the TPs by the sum of the TPs and the FPs (Equation 6).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \quad (6)$$

Empirical results

The study's findings were categorized into three groups: evaluating models trained on raw images, analyzing results with only background removal, and assessing the proposed preprocessing methods. By analyzing these different approaches, a comprehensive understanding of the effectiveness and performance of each method was obtained. Table I presents the results obtained from using raw images with CNN models. EfficientNet B0 showed slightly better performance than InceptionResNetV2, while RegNetX006 achieved the highest performance. RegNetX006 achieved the following metric values: accuracy (0.9775), F1 score (0.9725), MCC (0.9452), ROC-AUC (0.9706), sensitivity (0.9870), specificity (0.9542), and precision (0.9745).

Table II shows the results obtained using CNN models with preprocessed images where only

TABLE I
Performance comparison of CNN models on raw data classification

	Accuracy	F1 Score	MCC	ROC_AUC	Sensitivity	Specificity	Precision
RegNetX006	0.9775	0.9725	0.9452	0.9706	0.9870	0.9542	0.9745
EfficientNet B0	0.9399	0.9252	0.9452	0.9162	0.9724	0.8601	0.9357
InceptionResNetV2	0.9159	0.9036	0.8172	0.9310	0.8952	0.9667	0.8873

CNN: Convolutional neural network; MCC: Matthew's correlation coefficient; ROC-AUC: Receiver operating characteristic area under the curve.

TABLE II
Performance evaluation of CNN models on images only the background removed

	Accuracy	F1 Score	MCC	ROC_AUC	Sensitivity	Specificity	Precision
RegNetX006	0.9822	0.9821	0.9645	0.9824	0.9878	0.9770	0.9820
EfficientNet B0	0.9761	0.9761	0.9525	0.9765	0.9886	0.9644	0.9759
InceptionResNetV2	0.9391	0.9390	0.8781	0.9390	0.9407	0.9373	0.9389

CNN: Convolutional neural network; MCC: Matthew's correlation coefficient; ROC-AUC: Receiver operating characteristic area under the curve.

TABLE III
Performance analysis of CNN models on proposed preprocessed radiographic images

	Accuracy	F1 Score	MCC	ROC_AUC	Sensitivity	Specificity	Precision
RegNetX006	0.9921	0.9918	0.9842	0.9918	0.9974	0.9863	0.9923
EfficientNet B0	0.9725	0.9724	0.9452	0.9729	0.9622	0.9835	0.9722
InceptionResNetV2	0.9472	0.9490	0.8943	0.9287	0.9494	0.9448	0.9485

CNN: Convolutional neural network; MCC: Matthew's correlation coefficient; ROC-AUC: Receiver operating characteristic area under the curve.

the background was eliminated. EfficientNet B0 outperformed InceptionResNetV2, but RegNetX006 achieved the highest performance. RegNetX006 exhibited performance with an accuracy of 0.9822, an F1 score of 0.9821, a MCC of 0.9645, a ROC-AUC of 0.9824, a sensitivity of 0.9878, a specificity of 0.9770, and a precision of 0.9820.

Table III presents the results of RegNetX006, EfficientNet B0, and InceptionResNetV2 models with the proposed preprocessing method. EfficientNet B0 slightly outperformed InceptionResNetV2, while RegNetX006 demonstrated the most remarkable performance. By utilizing preprocessed radiographic images, RegNetX006 attained metric values, including accuracy (0.9921), F1 score (0.9918), MCC (0.9842), ROC-AUC (0.9918), sensitivity (0.9974), specificity (0.9863), and precision (0.9923).

DISCUSSION

The study compares the performance of three different deep learning models: RegNetX006, Efficient B0, and InceptionResNetV2. The results show that RegNetX006 consistently outperforms the other models in terms of accuracy, F1 score, MCC, ROC AUC, sensitivity, specificity, and precision. This indicates the superiority of RegNetX006 in detecting fractures of the ulna and radius.[14] Furthermore, the study evaluates the performance of the models using different types of input images: raw images, images with only the background removed, and preprocessed images. The highest classification performance is achieved when using the proposed preprocessing method in combination with the RegNetX006 model.[14]

In addition, performance improvement of RegNetX006 and InceptionResNetV2 models was observed at each step of preprocessing compared to raw data. The best classification performance achieved by the EfficientNet B0 model was carried out by using Images Only the Background Removed, and a relatively slight decrease in performance was observed in the last step. The reason for this situation can be explained as the fact that EfficientNet B0 has a simpler architecture and contains fewer hyperparameters compared to other models, it reaches saturation in the relevant step, and then the learning decreases.

There are several limitations to this study. Although the dataset used in the study includes different types of single or multiple fractures, it only consists of upper extremity radiography. Expanding the dataset and detecting different bone fractures is important to prove the effectiveness of the method. In future studies, the efficiency of the proposed method will be reinforced by expanding the data set on fractures in different bones. Nevertheless, the study contributes to the literature by its presentation of improved classification performance, detection of various fracture kinds, and architecture effectiveness: the suggested preprocessing method improves classification accuracy by removing noise; the study presents to detect numerous fractures across distinct types; next-generation architectures that have not yet been adopted by the literature perform well in fracture identification.

In conclusion, this article presented a method of detecting ulna and radius bone fractures using deep learning techniques in upper extremity radiographs. The findings show that using the proposed preprocessing strategy in combination with the RegNetX006 model provides the best classification performance.

Ethics Committee Approval: Since the dataset used in the study is open source in the literature, there was no need for an ethics committee approval.

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of Interest: The author declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding: The author received no financial support for the research and/or authorship of this article.

REFERENCES

1. Atik OŞ. Artificial intelligence, machine learning, and deep learning in orthopedic surgery. Jt Dis Relat Surg 2022;33:484-5. doi: 10.52312/jdrs.2022.57906.
2. Joshi D, Singh T. A survey of fracture detection techniques in bone X-ray images. Artificial Intelligence Review 2020;53:4475–517. doi: 10.1007/s10462-019-09799-0.
3. Vishnu V, Prakash J, Rengasamy S, Sharmila T. Detection and classification of long bone fractures. Int J Appl Eng Res 2015;10:18315-20.
4. Bayram, F, Çakiroglu M. DIFFRACT: Diaphyseal femur fracture classifier system. Biocybern Biomed Eng 2016;36:157-71.
5. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg 2020;31:175-83. doi: 10.5606/ehc.2020.72163.
6. Sahin ME. Image processing and machine learning-based bone fracture detection and classification using X-ray images. Int J Imaging Syst Technol 2023;33:853-65. doi: 10.1002/ima.22849.
7. Rashid T, Zia MS, Najam-Ur-Rehman, Meraj T, Rauf HT, Kadry S. A minority class balanced approach using the

DCNN-LSTM method to detect human wrist fracture. Life (Basel) 2023;13:133. doi: 10.3390/life13010133.
8. Bone Fracture Detection Using X-rays. Kaggle website. Available at: https://www.kaggle.com/datasets/vuppalaadithyasairam/bone-fracture-detection-using-xrays. [Accessed: 02.07.2023].
9. Bouwmans T, Javed S, Sultana M, Jung SK. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Netw 2019;117:8-66. doi: 10.1016/j.neunet.2019.04.024.
10. Baymurzina D, Golikov E, Burtsev M. A review of neural architecture search. Neurocomput 2022;474:82–93. doi: 10.1016/j.neucom.2021.12.014
11. Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, PMLR 2019;97:6105-14.
12. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17). San Francisco; 2017. p. 4278-84. doi: 10.1609/aaai.v31i1.11231
13. Erdaş Ç, Sumer E, Kibaroglu S. Neurodegenerative disease detection and severity prediction using deep learning approaches. Biomed Signal Process Control 2021;70:103069. doi: 10.1016/j.bspc.2021.103069.
14. Atik OŞ. Which articles do the editors prefer to publish? Jt Dis Relat Surg 2022;33:1-2. doi: 10.52312/jdrs.2022.57903.