ISSN 2075-3810 (Print), ISSN 2075-3829 (Online)

БІОФІЗИЧНИЙ ВІСНИК, 2026, Вип. 55 BIOPHYSICAL BULLETIN, 2026, Iss. 55

BIOMEDICAL ENGINEERING

Original article

https://doi.org/10.26565/2075-3810-2026-55-09

UDC 004.932.2:004.85:611.711

IMPROVING THE EFFICIENCY OF SPINE REGION SEGMENTATION USING AN ENSEMBLE OF PRE-TRAINED NEURAL NETWORKS

V. D. Koniukhov[1,*], O. M. Morgun[2], K. E. Nemchenko[3]

[1] National Scientific Center "Institute of Experimental and Clinical Veterinary Medicine", 83 Hryhoriia Skovorody St., Kharkiv, 61023, Ukraine;

[2] "Laboratory of X-ray Medical Equipment" LTD, 1 Dostoevsky St., Kharkiv, 61102, Ukraine;

[3] V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine

**Corresponding author: v.koniukhov.iecvm@gmail.com*

Submitted December 24, 2025; Revised May 18, 2026;
Accepted May 20, 2026; Published June 25, 2026

Background: The accuracy of segmentation of vertebrae in X-ray images is critical for clinical decisions as the manual method is laborious. The use of deep learning is complicated by low contrast, noise, and patient position artifacts. These negative factors make a single neural network unreliable. Thus, to improve the accuracy and efficiency of segmentation, regardless of the quality of X-ray images, there is a need for an ensemble of neural networks that compensates for the individual shortcomings of the models by aggregating their results.

Objectives: Increasing the accuracy and efficiency of segmentation of a spinal region consisting of four vertebrae (Th8, Th9, Th10, Th11) in X-ray images by using an ensemble of pre-trained neural networks.

Materials and methods: Two datasets were used for the experiments: the first set with 183 images was distributed in the ratio of 70% / 10% / 20% for training, validation, and testing, in turn, the second set of 58 images was used exclusively for the final assessment of the generalization ability of the ensemble on new data. In the process of research, segmentation accuracy with and without augmentation was first compared, after which the 10 best from the initial 20 neural networks were selected for further use, and five ensemble algorithms were used for mask aggregation.

Results: For the ensemble of pre-trained neural networks, the best result was shown by soft voting. Comparing the obtained result with the results presented by Koniukhov et al. (2024), the improvement was 3.06%. This indicator clearly confirms the effectiveness of using pre-trained networks for segmentation of the spine area.

Conclusions: Soft voting for an ensemble of pre-trained neural networks demonstrated the greatest improvement in segmentation accuracy compared to other methods. Aggregating knowledge from 10 models successfully eliminated the limitations of individual models. The use of an ensemble of pre-trained neural networks improved segmentation accuracy for both the test data from the first dataset and the data from the second dataset. Such results confirm the feasibility of applying the proposed ensemble-based approach to chest X-ray radiographs for vertebrae segmentation in medical imaging tasks.

KEY WORDS: image segmentation; deep learning; ensemble learning; medical imaging; neural networks; spinal diseases.

Spinal diseases are a common problem in our time. Such diseases as: osteochondrosis, spondylosis, spondylarthrosis, deforming arthrosis, scoliosis, kyphosis, lordosis and others are

Citation: Koniukhov VD, Morgun OM, Nemchenko KE. Improving the efficiency of spine region segmentation using an ensemble of pre-trained neural networks. Biophysical Bulletin. 2026;55:117–129. https://doi.org/10.26565/2075-3810-2026-55-09

a global problem and lead to significant pain for a patient. For example, low back pain alone affected 619 million people in 2020 and by 2050, 843 million are predicted [1]. Timely diagnosis of diseases is a key necessity for building effective treatment. The use of medical imaging methods is an integral part of the diagnostic process. Medical imaging can also be used to monitor and plan treatment. There are a significant number of medical imaging methods: radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound diagnostics. X-ray images are the cheapest and most accessible, as well as a powerful tool for obtaining a primary diagnosis [2]. Although radiography is often underestimated in comparison with such methods as MRI or CT, using X-ray images doctors obtain the necessary information [3].

Manual radiograph analysis is too often subject to variability. For example, a study [4] supports this idea, based on high inter- and intra-operator error in calculating the Cobb angle. Such subjectivity can negatively affect the choice of treatment tactics and disease monitoring. To solve this problem, it is recommended to use automatic methods, which are considered more reliable [5]. Similar problems are also observed when performing other tasks associated with the manual method.

Deep learning methods can be used to solve problems associated with manual methods. It is deep learning models that have taken image segmentation to a new level compared to traditional methods [6]. Every year, deep learning methods demonstrate better results, as not only their quality improves, but also the number of publicly available data sets for training increases, which significantly accelerates research in this area [7]. The use of deep learning methods has made it possible to significantly increase the accuracy and efficiency of image analysis and successfully perform the following tasks: classification of chest X-ray images [8]; detection of COVID-19 [9]; recognition of periodontitis and dental caries on X-ray images [10]; detection of wrist fractures on X-ray images [11]; detection and classification of knee osteoarthritis [12]; detection of grain defects on X-ray images [13]; detection of anomalies on shoulder X-ray images [14]; detection of vertebral fractures on X-ray images [15]. The listed works are a small tip of the iceberg in the use of deep learning methods. It is obvious that when any new high-performance methods appear, scientists always try to apply the corresponding methods in their applied problems. Such improvements include not only deep learning methods but also various ensemble approaches or multi-stage methods. Using a multi-stage deep learning method, it is possible to first find a region of the spine and then isolate the necessary vertebrae that are in this section [16]. Using ensemble methods, it is possible to aggregate the predictions of several models, reducing the impact of random errors of individual models on the final mask [17]. Architectures such as U-Net [18] can use powerful backbones (e.g. ResNet50, SEResNext50, EfficientNetB0) to extract high-level features. Thus, we have the opportunity to use the weights of a model that has already been pre-trained on a large dataset. This approach allows the model to quickly identify both low-level and medium-level features that are relevant to natural and medical images. In our previous work [19], using an ensemble approach for spine segmentation, we were able to obtain a Dice-Sørensen coefficient (DSC) of 81.79% for an independent dataset. In order to increase the reliability of the results, this study proposes to use an ensemble of pre-trained neural networks and some other improvements.

MATERIALS AND METHODS

Datasets, preprocessing, and evaluation metrics

Open-source chest X-ray images of human patients were used for training, validation, and testing. The first dataset [20] consists of 183 radiographic images, which were split into 70% for training, 10% for validation, and 20% for testing. Using the second resource [21], an

additional test dataset of 58 images was created. The thoracic vertebrae Th8, Th9, Th10, and Th11 were selected as the target anatomical structures for the segmentation task. All images were resized to the same size of 512×512 pixels and converted to grayscale. Segmentation masks for all images in both datasets were created by O. M. Morgun.

No additional methods were used at the preprocessing stage, which allowed us to preserve the original data structure.

Seven metrics were used to evaluate the performance: Intersection over Union (IoU), DSC, recall, specificity, precision, F1-score, pixel accuracy (PA). The corresponding formulas are given below:

$$IoU = \frac{TP}{TP + FP + FN}. \quad (1)$$

IoU is a metric used to evaluate the overlap between a predicted and a ground truth.

$$DSC = \frac{2TP}{2TP + FP + FN}. \quad (2)$$

DSC is a standard metric used to evaluate the similarity between a predicted mask and a corresponding ground truth mask by measuring the amount overlap between them.

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

Recall represents an ability of a model to correctly identify all relevant cases.

$$Specificity = \frac{TN}{TN + FP}. \quad (4)$$

Specificity measures a proportion of true negative cases that are correctly identified by a model.

$$Precision = \frac{TP}{TP + FP}. \quad (5)$$

Precision shows a proportion of positive identifications that were actually correct.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (6)$$

F1-score is a harmonic mean of precision and recall.

$$PA = \frac{TP + TN}{TP + TN + FP + FN}. \quad (7)$$

Pixel accuracy calculates a ratio of all correctly classified pixels to the total number of pixels, where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

Computational complexity

The computational efficiency of the ensemble was evaluated on a workstation running Windows 11 Pro and equipped with an Intel Core i7-10700, NVIDIA GeForce RTX 4060 Ti (16GB VRAM), and 32 GB of RAM. To measure the inference time, we calculated the average duration of a single forward pass for a 512×512 image through the entire ensemble of 10 models. Memory consumption was also monitored at the prediction stage.

Stage 1: Comparative analysis of augmentation robustness

The first stage of the research compared two approaches – with and without augmentation. U-Net was used as the basic neural network architecture. Both options were trained on the same datasets and with the same hyperparameters. The main goal was to demonstrate the benefits of fine-tuning the augmentation parameters, which can potentially improve the overall performance of the model. For each approach, the model was trained 50 times, after which the average values of the corresponding metrics were calculated.

All input images were resized to a uniform size of 256×256 pixels. The training process was performed using a batch size of 4, the *Adam* optimizer, a learning rate of 1×10^{-4}, and a *binary cross-entropy* loss function. To prevent overfitting an *early stopping* mechanism with *patience*=10 and *validation* DSC monitoring were used, while the best model weights were kept. Each model was trained for a maximum of 100 epochs.

Augmentation parameters included random image rotation from −10° to 10°, scaling from 0.8 to 1.2, shift from ±10% of the image size, and brightness change from −20 to +20 units. In addition, pixel intensities could be multiplied by a factor of 0.7 to 1.3, Gaussian blurring with σ in the range of 0.0-0.4 was applied, and with a probability of 30% the image remained unchanged.

Stage 2: Statistical selection of ensemble candidate architectures

In the second stage, the U-Net model was trained using twenty different backbones: None, ResNet34, ResNet50, ResNet152, SEResNet34, SEResNet50, SEResNet152, ResNeXt50, SEResNeXt50, DenseNet121, InceptionV3, InceptionResNetV2, MobileNetV2, EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6. For each backbone, the training process was repeated 10 times, which allowed us to average the results, reduce the influence of random factors, and assess the stability of performance indicators. After a successful training process, the top 10 best models were selected based on their performance on the validation subset of the first dataset. The selection criterion was the DSC ≥ 91.5% on the first dataset. These ten models will subsequently be used in the third stage. At this stage all input images were resized to a same size of 256×256 pixels. During training a batch size of 4 was used and the maximum number of epochs was 300. ImageNet was used as pre-trained weights for the encoder, which provided better initial generalizability of the model. The optimization process was carried out using the *Adam* optimizer with a learning rate of 1×10^{-4}, and *binary cross-entropy* was chosen as the loss function. To prevent overtraining, an *early stopping* mechanism was used, which monitored the value of the validation DSC metric in *max* mode with the *patience*=25 parameter. If there was no improvement in the metric within the specified number of epochs, training was stopped and the model restored the best weights. The augmentation parameters were used from the first stage.

Stage 3: Final training of base ensemble models

At this stage, the ensemble of neural networks was trained. Ten models from the second stage were selected, with the following backbones: SEResNet50, SEResNet152,

SEResNeXt50, DenseNet121, InceptionV3, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB6. Almost all the parameters specified for the second stage were used for training, except for the following – all images were resized to the same size of 512×512 pixels, and the training rate was also changed to 1×10^{-5}.

Stage 4: Comparative ensemble combination and validation

At the final stage, a comparison of five ensemble methods was conducted: the max voting method, the min voting method, the soft voting method, the shape averaging method, and the union ensemble method. In all methods where a threshold *T* was used, we set *T*=0.5. All of these methods were applied to the ensemble defined in the third point, with the aim of identifying the most effective mask aggregation method. For the mathematical description of the ensemble methods, let $P_k(i,j)$ be the probability of pixel (i,j) belonging to the target class, as predicted by the *k*-th model, where *M* is the total number of models in the ensemble.

Max voting. In this method, the final ensemble decision is based on the maximum probability predicted across all models for each specific pixel. The ensemble $E_{max}(i,j)$ result is defined as:

$$P_{max}(i,j) = \max_{k\in\{1,\dots,M\}} P_k(i,j)\,. \tag{8}$$

$$E_{max}(i,j) = \begin{cases} 1, & if\ P_{max}(i,j) > T \\ 0, & otherwise \end{cases}. \tag{9}$$

Min voting. In this method, the ensemble predicts the target class only if even the minimum probability across all models exceeds the binarization threshold. Then the resulting ensemble probability $P_{min}(i,j)$ and the binary prediction $E_{min}(i,j)$ are defined as:

$$P_{min}(i,j) = \min_{k\in\{1,\dots,M\}} P_k(i,j)\,. \tag{10}$$

$$E_{min}(i,j) = \begin{cases} 1, & if\ P_{min}(i,j) > T \\ 0, & otherwise \end{cases}. \tag{11}$$

Soft voting. This method is based on calculating the arithmetic mean of the prediction values from all ensemble models for each pixel, followed by binarization of the obtained average. The average probability value $\bar{P}(i,j)$ and the final binary prediction $E_{soft}(i,j)$ are determined by the following formulas:

$$\bar{P}(i,j) = \frac{1}{M}\sum_{k=1}^{M} P_k(i,j)\,. \tag{12}$$

$$E_{soft}(i,j) = \begin{cases} 1, & if\ \bar{P}(i,j) > T \\ 0, & otherwise \end{cases}. \tag{13}$$

Union ensemble. This method is based on the logical addition (OR) operation. This means that a pixel is considered part of the target object if it has been classified as such by at least one model in the ensemble. Let S_k be a binary mask obtained from the *k*-th ensemble model. The final combined mask E_{union} is determined through the set union operation:

$$E_{union} = \bigcup_{k=1}^{M} S_k \,. \tag{14}$$

Shape averaging. This method is based on aggregating the geometric properties of the masks by transforming them into signed distance fields. For each binary mask S_k, a signed distance matrix D_k is calculated as the difference between the distance transform of the mask and its inverted mask. The individual matrices are then accumulated into a global distance matrix $\overline{D}$ through an element-wise summation of al ensemble models. The final binary ensemble prediction E_{shape} is obtained by selecting pixels where the total accumulated distance value is positive. For each mask, a matrix of signed distances D_k is calculated:

$$D_k(i,j) = DT(S_k) - DT(\neg S_k), \tag{15}$$

where DT is a distance transformation function that calculates the distances from pixel (i,j) to the nearest boundary of the object. The value $D_k(i,j)$ is positive for pixels inside the object and negative for background pixels.
The aggregated distance matrix $\overline{D}$ is calculated as the sum of individual transformations:

$$\overline{D}(i,j) = \sum_{k=1}^{M} D_k(i,j) \,. \tag{16}$$

The final binary ensemble mask $E_{shape}(i,j)$ is obtained by selecting pixels with a positive total distance value:

$$E_{shape}(i,j) = \begin{cases} 1, & if\ \overline{D}(i,j) > 0 \\ 0, & otherwise \end{cases} . \tag{17}$$

RESULTS AND DISCUSSION

To illustrate the X-ray data used to train and test the models, example X-ray images and corresponding masks are shown in Figure 1. The top row shows an example image from the first set and the bottom row shows an image from the second set.

We acknowledge that the segmentation masks were created by a single author. Although this ensured consistency in the annotation style across images, it introduces potential annotation bias and does not allow for analysis of interobserver variability. This limitation will be addressed in future work by involving multiple independent radiologists.

At the beginning of the research, it was proposed to consider the impact of data augmentation on its generalization ability. Augmentation parameters were selected in such a way as to simulate natural variations in biomedical images as accurately as possible.

The results are shown in Table 1 demonstrating the improvement of the DSC for the variant that used data augmentation. The DSC increased by 2.53% and the IoU by 3.85%. The noticeable increase in efficiency confirms the fact that the correct setting of the augmentation parameters can be used as a factor to counteract overfitting and as a general factor to increase the generalization ability. In the future, all stages will be performed using the described set of data augmentation parameters.

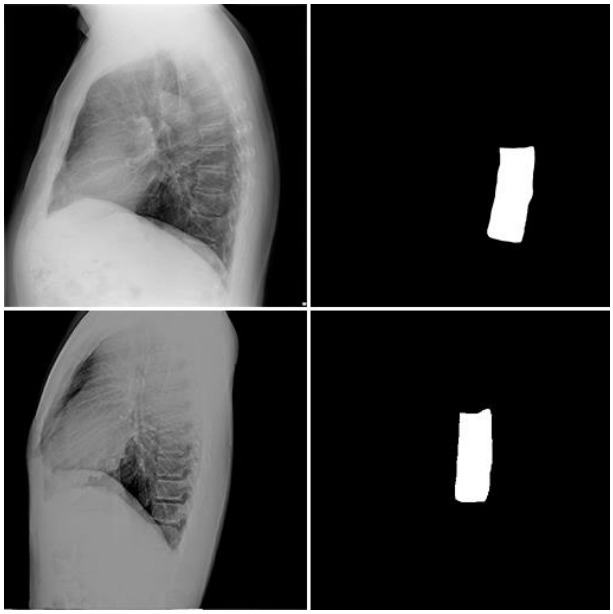

Fig. 1. Examples of chest X-ray radiographs of human patients and their corresponding ground truth segmentation masks of thoracic vertebrae, illustrating differences in image quality between the datasets. The top row shows a sample from dataset 1 [20], while the bottom row shows a sample from dataset 2 [21].

Table 1. Comparison of U-Net with VGG16 backbone performance metrics (%): data augmentation vs. no augmentation (dataset 1)

Type	IoU	DSC	Recall	Specificity	Precision	F1	PA
With augmentation	86.61	92.57	90.56	99.81	95.04	92.57	99.46
Without augmentation	82.76	90.04	91.90	99.48	89.11	90.04	99.19

The number and quality of models in the ensemble play a key role. To improve the generalization ability, it is necessary to include backbones of different architectures in the ensemble. At the second stage a comparison of the efficiency of U-Net using different backbones was performed. The results shown in Tables 2, 3 made it possible to determine the optimal backbones for performing the task. The poor results obtained for the backbones: ResNet34, ResNet50, ResNet152, SEResNet, ResNeXt50 may indicate insufficient relevance for performing a specific task. Such a result could also be influenced by a too sharp decrease in spatial resolution in the initial layers. Comparing the results of SEResNet and ResNet, one can see a huge difference in efficiency. Such a result is most likely due to the use of attention mechanisms in SEResNet. The comparative analysis of 20 backbones allowed for the selection of the optimal top ten architectures that demonstrated the highest accuracy on the first dataset, with DSC scores exceeding 91.5% (Table 2). These selected models were then evaluated using the second dataset to test their generalization ability. A performance drop averaging 7.61% was observed on the second dataset due to lower image quality (Table 3). This is expressed in lower contrast or higher noise levels.

Table 2. Comparison of U-Net segmentation performance metrics (%) across different backbone architectures (dataset 1)

Backbone	IoU	DSC	Recall	Specificity	Precision	F1	PA
None	84.85	91.53	89.46	99.78	94.15	91.53	99.38
ResNet34	8.34	14.79	73.82	57.07	18.68	14.79	57.69
ResNet50	7.29	12.95	65.99	65.34	7.55	37.79	65.36
ResNet152	6.37	11.13	62.63	57.79	6.84	38.93	57.98
SEResNet34	8.43	13.85	39.46	79.67	14.61	41.30	78.17
SEResNet50	87.19	92.93	91.70	99.78	94.61	92.93	99.47
SEResNet152	87.36	93.07	91.58	99.80	95.00	93.07	99.49
ResNeXt50	9.16	15.59	81.18	51.53	19.60	15.86	52.62
SEResNeXt50	86.77	92.69	90.61	99.81	95.31	92.69	99.46
DenseNet121	86.60	92.60	91.30	99.77	94.22	92.60	99.44
InceptionV3	86.55	92.56	90.28	99.82	95.45	92.56	99.46
InceptionResNetV2	84.63	91.49	89.31	99.78	94.24	91.49	99.38
MobileNetV2	73.37	83.95	77.68	99.76	93.13	83.95	98.93
EfficientNetB0	83.27	90.53	88.27	99.75	93.52	90.53	99.31
EfficientNetB1	85.81	92.09	90.36	99.77	94.24	92.09	99.42
EfficientNetB2	85.46	91.86	90.24	99.76	93.97	91.86	99.39
EfficientNetB3	85.99	92.21	89.67	99.83	95.40	92.21	99.44
EfficientNetB4	84.98	91.58	89.57	99.77	94.03	91.58	99.38
EfficientNetB5	84.57	91.27	89.89	99.72	93.29	91.27	99.35
EfficientNetB6	85.53	91.91	89.68	99.79	94.58	91.91	99.40

Table 3. Comparison of U-Net segmentation performance metrics (%) across different backbone architectures (dataset 2)

Backbone	IoU	DSC	Recall	Specificity	Precision	F1	PA
None	70.39	81.50	81.13	99.42	82.72	81.50	98.80
ResNet34	7.27	12.98	73.11	56.28	17.53	13.32	56.83
ResNet50	7.22	12.84	67.68	64.96	7.98	34.20	65.04
ResNet152	5.16	9.18	59.54	58.23	5.44	39.95	58.29
SEResNet34	4.66	7.95	29.30	79.73	8.43	53.17	78.08
SEResNet50	73.61	83.76	83.27	99.50	85.13	83.76	98.95
SEResNet152	73.96	84.02	83.51	99.51	85.65	84.02	98.96
ResNeXt50	8.09	13.86	81.14	50.97	18.60	14.72	51.98
SEResNeXt50	73.45	83.65	81.45	99.58	87.18	83.65	98.97
DenseNet121	73.58	83.82	84.08	99.45	84.27	83.82	98.93
InceptionV3	72.65	83.14	82.72	99.46	84.31	83.14	98.89
InceptionResNetV2	70.90	81.31	80.72	99.45	83.55	82.26	98.80
MobileNetV2	57.07	69.39	61.55	99.74	90.54	69.73	98.50
EfficientNetB0	70.05	80.81	80.06	99.44	82.87	82.19	98.79
EfficientNetB1	72.82	83.31	82.75	99.48	84.72	83.31	98.90
EfficientNetB2	73.23	83.51	83.30	99.47	84.61	83.51	98.92
EfficientNetB3	73.37	83.57	83.65	99.46	84.05	83.57	98.92
EfficientNetB4	72.92	83.28	83.43	99.46	84.36	83.28	98.91
EfficientNetB5	72.45	82.97	83.70	99.39	83.31	82.97	98.85
EfficientNetB6	73.24	83.49	83.04	99.48	84.78	83.49	98.91

Tables 4 and 5 show the training results of the 10 best backbones. The key differences in the training process in the second and third stages were the number of training sessions for each model and the size of the images, as well as the change in the training speed. If in the second stage training was performed 10 times for each model to select the best models, then in the third stage, training was performed only once. Another significant change was the use of different image sizes, in the second stage images of 256×256 pixels were used and in the third 512×512 pixels. This difference is due to the need to perform statistical analysis in the second stage, since it was necessary to train the model 10 times. Reducing the size made it possible to reduce the number of calculations and memory consumption.

Based on the results presented in Tables 4 and 5, statistical indicators (min, max, mean) were calculated. These data are given in Tables 6 and 7, respectively. Analysis of the obtained results showed that there is a gap in the generalization ability of the models between the two datasets. The difference in the average DSC between the first and second datasets was 9.66%. Whereas, when comparing 20 models, a DSC of 7.61% was obtained. Comparison of the results obtained at the third stage with the results of the second stage also demonstrates positive dynamics: the maximum value increased by 1.99% for the first set and by 0.5% for the second. In general, all backbone architectures at the third stage demonstrated high efficiency indicators, both for the first set and for the second. Final training was carried out for 10 U-Net models, each of which used a unique backbone architecture. The main goal was to create a diversified ensemble, because models with different backbones may focus on different features of X-ray images.

Table 4. Comparison of segmentation performance metrics (%) for top 10 models (dataset 1)

Backbone	IoU	DSC	Recall	Specificity	Precision	F1	PA
SEResNet50	90.83	95.06	95.74	99.77	94.71	95.06	99.62
SEResNet152	88.47	93.69	93.72	99.75	93.98	93.69	99.52
SEResNeXt50	87.52	93.17	93.71	99.70	93.18	93.17	99.47
DenseNet121	87.54	93.08	93.78	99.71	92.84	93.08	99.47
InceptionV3	89.10	94.04	94.42	99.75	94.03	94.04	99.55
EfficientNetB1	87.26	92.89	93.04	99.72	93.38	92.89	99.47
EfficientNetB2	87.70	93.08	92.05	99.78	94.49	93.08	99.48
EfficientNetB3	86.64	92.53	91.79	99.75	93.79	92.53	99.45
EfficientNetB4	87.52	93.04	91.82	99.79	94.76	93.04	99.48
EfficientNetB6	85.93	92.10	92.20	99.69	92.55	92.10	99.41

Table 5. Comparison of segmentation performance metrics (%) for top 10 models (dataset 2)

Backbone	IoU	DSC	Recall	Specificity	Precision	F1	PA
SEResNet50	73.13	83.55	87.37	99.28	80.74	83.55	98.89
SEResNet152	73.03	83.45	85.96	99.33	81.68	83.45	98.88
SEResNeXt50	72.16	82.89	86.14	99.28	80.72	82.89	98.84
DenseNet121	73.67	83.85	86.29	99.36	82.23	83.85	98.93
InceptionV3	73.13	83.31	84.85	99.38	83.12	83.31	98.92
EfficientNetB1	73.40	83.68	86.03	99.36	82.04	83.68	98.92
EfficientNetB2	74.12	84.19	85.09	99.45	84.34	84.19	98.98
EfficientNetB3	72.78	83.30	82.86	99.49	84.48	83.30	98.93
EfficientNetB4	74.82	84.51	85.07	99.47	84.58	84.51	98.99
EfficientNetB6	72.60	83.33	84.68	99.41	82.94	83.33	98.92

Table 6. Statistical summary of segmentation performance metrics (%) for top 10 models (dataset 1)

Statistic	IoU	DSC	Recall	Specificity	Precision	F1	PA
Min	85.93	92.10	91.79	99.69	92.55	92.10	99.41
Max	90.83	95.06	95.74	99.79	94.76	95.06	99.62
Mean	87.85	93.27	93.23	99.74	93.77	93.27	99.49

Table 7. Statistical summary of segmentation performance metrics (%) for top 10 models (dataset 2)

Statistic	IoU	DSC	Recall	Specificity	Precision	F1	PA
Min	72.16	82.89	82.86	99.28	80.72	82.89	98.84
Max	74.82	84.51	87.37	99.49	84.58	84.51	98.99
Mean	73.28	83.61	85.44	99.38	82.69	83.61	98.92

At the final stage, the effectiveness of various aggregation methods for creating an ensemble was assessed. The goal was to determine the optimal strategy for combining individual predictions to achieve maximum accuracy. Analysis of Tables 8 and 9 demonstrates that the soft voting method provides the highest efficiency, although the shape averaging method is a close second. The success of the soft voting method is associated with effective error smoothing and its ability to compensate for uncertainty at the boundaries of objects. Having achieved a DSC of 94.21% for the first set and 84.85% for the second set, the soft voting method demonstrates that it is the optimal ensemble technique, since it is the one that best uses the diversification of features extracted by different backbones.

Table 8. Comparison of segmentation performance metrics (%) across different ensemble methods (dataset 1)

Ensemble method	IoU	DSC	Recall	Specificity	Precision	F1	PA
Min voting	85.31	91.77	86.31	99.95	98.68	91.77	99.43
Max voting	82.48	90.24	98.15	99.23	83.87	90.24	99.19
Soft voting	89.42	94.21	93.20	99.83	95.73	94.21	99.57
Shape averaging	89.33	94.17	93.32	99.82	95.44	94.17	99.57
Union ensemble	82.61	90.30	98.15	99.23	84.00	90.30	99.19

Analysis of the ensemble's operational performance revealed that the average inference time per single X-ray image for the entire 10-model ensemble is 226 ms. While this is slower than a single model (which averages 22-26 ms), the speed of the ensemble remains effective for clinical diagnostic tasks where real-time processing is not strictly necessary. The peak GPU memory usage during the inference of the most complex backbone did not exceed 6 GB of VRAM.

Table 9. Comparison of segmentation performance metrics (%) across different ensemble methods (dataset 2)

Ensemble method	IoU	DSC	Recall	Specificity	Precision	F1	PA
Min voting	71.38	82.07	76.22	99.74	91.21	82.07	98.97
Max voting	66.95	79.39	90.41	98.74	71.20	79.39	98.46
Soft voting	75.15	84.85	85.93	99.46	84.38	84.85	99.01
Shape averaging	74.78	84.66	85.55	99.46	84.44	84.66	99.00
Union ensemble	67.51	79.85	90.24	98.80	72.04	79.85	98.51

To better position our contribution within the current state-of-the-art, we compared our results with several recent studies on spine segmentation across different imaging modalities. Qadri et al. (2022) [22] utilized stacked sparse autoencoders for patch-based classification in CT images (SVseg), achieving DSC of 87.39%; in comparison, our ensemble approach on X-ray images demonstrated performance with DSC of 84.85%. Similarly, Lu et al. (2023) [23] proposed a two-stage localization and segmentation framework (XUnet) for CT scans. Their method requires a separate localization step to handle the complexity of lumbar vertebrae. In contrast, our method achieved high accuracy by leveraging the diversity of 10 pre-trained models without the need for a separate localization network. Furthermore, van der Graaf et al. (2024) [24] emphasized the effectiveness of nnU-Net for 3D volumetric data. Our study confirms that an ensemble with soft voting and shape averaging is equally robust for 2D X-ray radiography. Our method effectively compensates for lower contrast and overlapping structures, providing competitive anatomical consistency relative to these recent state-of-the-art frameworks.

CONCLUSIONS

Correctly setting the data augmentation parameters provided an increase in the DSC by 2.53% compared to the option without data augmentation. Ensemble deep learning methods represent an effective approach to enhance the accuracy of X-ray image segmentation. The results suggest that ensembles leveraging models with diverse architectures or backbones can effectively combine the strengths of individual models, potentially reducing the impact of specific errors. In this study, an ensemble of the 10 best-performing models was found to be an effective compromise for improving segmentation accuracy.

A comparison of the main ensemble mask aggregation methods has shown that for segmentation tasks with high requirements (as in the case of X-ray images), the use of the soft voting method with threshold 0.5 is the most efficient option. While the shape averaging method showed slightly different morphological results, it remains a valid alternative. The comparison with previous work [19] showed an improvement in the DSC from 81.79% to 84.85% for the second dataset. This 3.06% increase highlights the benefits of using pre-trained backbones in biomedical segmentation tasks.

The obtained results may be useful for improving computer-aided analysis of chest X-ray images. It may also support more accurate vertebrae segmentation in medical imaging workflows, contributing to the development of decision-support tools for clinical diagnostics.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

Authors' ORCID ID

V. D. Koniukhov https://orcid.org/0009-0007-0256-1388
O. M. Morgun https://orcid.org/0009-0005-6157-9110
K. E. Nemchenko https://orcid.org/0000-0002-0734-942X

REFERENCES

1. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol. 2023;5(6):e316–e329. https://doi.org/10.1016/S2665-9913(23)00098-X
2. Ou X, Chen X, Xu X, Xie L, Chen X, Hong Z, et al. Recent development in X-ray imaging technology: future and challenges. Research. 2021;2021:9892152. https://doi.org/10.34133/2021/9892152
3. Goodwin ML, Buchowski JM, Sciubba DM. Why X-rays? The importance of radiographs in spine surgery. Spine J. 2022;22(11):1759–67. https://doi.org/10.1016/j.spinee.2022.07.102
4. Jin C, Wang S, Yang G, Li E, Liang Z. A review of the methods on Cobb angle measurements for spinal curvature. Sensors. 2022;22(9):3258. https://doi.org/10.3390/s22093258
5. Wills BP, Auerbach JD, Zhu X, Caird MS, Horn BD, Flynn JM, et al. Comparison of Cobb angle measurement of scoliosis radiographs with preselected end vertebrae: traditional versus digital acquisition. Spine. 2007;32(1):98–105. https://doi.org/10.1097/01.brs.0000251086.84420.d1
6. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. IEEE Trans. Pattern Anal Mach Intell. 2022;44(7):3523–42. https://doi.org/10.1109/TPAMI.2021.3059968
7. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K. Deep learning for chest X-ray analysis: a survey. Med Image Anal. 2021;72:102125. https://doi.org/10.1016/j.media.2021.102125
8. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. Sci Rep. 2019;9(1):6381. https://doi.org/10.1038/s41598-019-42294-8
9. Akter S, Shamrat F, Chakraborty S, Karim A, Azam S. COVID-19 detection using deep learning algorithm on chest X-ray images. Biology. 2021;10(11):1174. https://doi.org/10.3390/biology10111174
10. Chen IDS, Yang C-M, Chen M-J, Chen M-C, Weng R-M, Yeh C-H. Deep learning-based recognition of periodontitis and dental caries in dental X-ray images. Bioengineering. 2023;10(8):911. https://doi.org/10.3390/bioengineering10080911
11. Hardalaç F, Uysal F, Peker O, Çiçeklidağ M, Tolunay T, Tokgöz N, et al. Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors. 2022;22(3):1285. https://doi.org/10.3390/s22031285
12. Abdullah SS, Rajasekaran MP. Automatic detection and classification of knee osteoarthritis using deep learning approach. Radiol med. 2022;127(3):398–406. https://doi.org/10.1007/s11547-022-01476-7
13. Hamdy S, Charrier A, Corre LL, Rasti P, Rousseau D. Toward robust and high-throughput detection of seed defects in X-ray images via deep learning. Plant Methods. 2024;20(1):63. https://doi.org/10.1186/s13007-024-01195-2
14. Alzubaidi L, Salhi A, Fadhel MA, Bai J, Hollman F, Italia K, et al. Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images. PLoS One. 2024;19(3):e0299545. https://doi.org/10.1371/journal.pone.0299545
15. Cheng L-W, Chou H-H, Cai Y-X, Huang K-Y, Hsieh C-C, Chu P-L, et al. Automated detection of vertebral fractures from X-ray images: A novel machine learning model and survey of the field. Neurocomputing. 2024;566:126946. https://doi.org/10.1016/j.neucom.2023.126946
16. Koniukhov V. Improving the segmentation of the vertebrae using a multi-stage machine learning algorithm. Radioelectronic and Computer Systems. 2024;(4):83–90. https://doi.org/10.32620/reks.2024.4.07
17. Koniukhov VD, Morgun OM, Nemchenko KE. Impact of preprocessing and comparison of neural network ensemble methods for segmentation of the thoracic spine in X-ray images. Radio Electronics, Computer Science, Control. 2024;4:102–12. https://doi.org/10.15588/1607-3274-2024-4-10
18. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science, vol 9351. Cham: Springer; 2015. p. 234-41. https://doi.org/10.1007/978-3-319-24574-4_28
19. Koniukhov VD, Morgun OM, Nemchenko KE. Comparative analysis of ensemble methods for X-ray images: study of the influence on trained and new data. In: Computer modeling in science-intensive technologies:

proceedings of the International scientific and technical conference. Kharkiv; 2024. p. 100–103. (In Ukrainian). Available from: https://drive.google.com/file/d/1Vyvntnwud71YE21aGqyZMb483fAQL6pe/
20. Stanford AIMI. CheXpert Chest X-Rays [Internet]. AIMI Center. [cited 2025 Nov 16]. Available from: https://aimi.stanford.edu/datasets/chexpert-chest-x-rays
21. Vindr.ai. Vindr.ai datasets: SpineXR [Internet]. Vindr.ai. [cited 2025 Nov 16]. Available from: https://vindr.ai/spinexr
22. Qadri SF, Shen L, Ahmad M, Qadri S, Zareen SS, Akbar MA. SVseg: stacked sparse autoencoder-based patch classification modeling for vertebrae segmentation. Mathematics. 2022;10(5):796. https://doi.org/10.3390/math10050796
23. Lu H, Li M, Yu K, Zhang Y, Yu L. Lumbar spine segmentation method based on deep learning. J Appl Clin Med Phys. 2023;24(6):e13996. https://doi.org/10.1002/acm2.13996
24. van der Graaf JW, van Hooff ML, Buckens CFM, Rutten M, van Susante JLC, Kroeze RJ, et al. Lumbar spine segmentation in MR images: a dataset and a public benchmark. Scientific Data. 2024;11(1):264. https://doi.org/10.1038/s41597-024-03090-w

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ СЕГМЕНТАЦІЇ ОБЛАСТІ ХРЕБТА ЗА ДОПОМОГОЮ АНСАМБЛЮ ПОПЕРЕДНЬО НАВЧЕНИХ НЕЙРОННИХ МЕРЕЖ

В. Д. Конюхов [1], О. М. Моргун [2], К. Е. Нємченко [3]

[1] Національний науковий центр «Інститут експериментальної і клінічної ветеринарної медицини», вул. Григорія Сковороди, 83, м. Харків, 61023, Україна;
[2] ТОВ «Лабораторія рентгенівської медичної техніки», вул. Достоєвського, 1, м. Харків, 61102, Україна;
[3] Харківський національний університет імені В. Н. Каразіна, майдан Свободи, 4, м. Харків, 61022, Україна
e-mail: v.koniukhov.iecvm@gmail.com

Надійшла до редакції 24 грудня 2025 р. Переглянута 18 травня 2026 р.
Прийнята до друку 20 травня 2026 р. Опублікована 25 червня 2026 р.

Актуальність. Точність сегментації хребців на рентгенівських знімках є критичною для клінічних рішень, оскільки ручний метод є трудомістким. Використання глибокого навчання ускладнене через низький контраст, шум та артефакти положення пацієнта. Ці негативні фактори роблять одну нейронну мережу ненадійною. Таким чином, для покращення точності та ефективності сегментації, незалежно від якості рентгенівських знімків, виникає потреба в ансамблі нейронних мереж, який нівелює індивідуальні недоліки моделей шляхом агрегації їхніх результатів.
Мета роботи — підвищення точності та ефективності сегментації ділянки хребта, яка складається із чотирьох хребців (Th8, Th9, Th10, Th11), на рентгенівських знімках завдяки використанню ансамблю попередньо навчених нейронних мереж.
Матеріали і методи. Для проведення експериментів було використано два набори даних: перший набір зі 183 зображеннями було розподілено у співвідношенні 70% / 10% / 20% для навчання, валідації та тестування, в свою чергу другий набір із 58 зображень застосовувався виключно для фінальної оцінки генералізаційної здатності ансамблю на нових даних. У процесі дослідження спочатку порівнювали точність сегментації з аугментацією та без, після чого з початкових 20 моделей було відібрано 10 найкращих для подальшого використання, а для агрегації масок було використано п’ять ансамблевих алгоритмів.
Результати. Для ансамблю попередньо навчених нейронних мереж найкращий результат показав метод м’якого голосування. Порівняння отриманого результату з даними, наведеними у праці Конюхова В. Д. та ін. (2024), продемонструвало покращення на 3.06%. Такий показник однозначно підтверджує ефективність використання попередньо навчених мереж для сегментації ділянки хребта.
Висновки. Метод м’якого голосування для ансамблю попередньо навчених нейронних мереж продемонстрував найбільше покращення точності сегментації в порівнянні з іншими методами. Агрегування знань із 10 моделей успішно нівелювало недоліки використання кожної з них окремо. Використання ансамблю попередньо навчених нейронних мереж покращило точність сегментації як для тестових даних першого набору, так і для другого набору даних. Такі результати підтверджують доцільність застосування запропонованого ансамблевого підходу для сегментації хребців на рентгенівських зображеннях органів грудної клітки в медичних задачах.
КЛЮЧОВІ СЛОВА: сегментація зображень; глибоке навчання; ансамблеве навчання; медична візуалізація; нейронні мережі; захворювання хребта.