Exploiting advanced data augmentation strategies to improve the classification of spinal disorders in X-ray images

Omitted due to the double-blind review

***Abstract*—Spinal disorders affect a significant portion of the population and are a growing concern for healthcare authorities due to their potentially debilitating consequences. In this context, computer vision techniques offer a promising path for rapid diagnosis, but they often require large datasets to train robust and reliable models. However, public datasets that can support the training of such models are scarce. To address this, we explored the application of advanced data augmentation strategies, namely CutMix, CutOut, and MixUp, combined with standard augmentation techniques to improve the performance of deep learning models in classifying X-ray images into three categories: (i) healthy, (ii) scoliosis, and (iii) spondylolisthesis. We applied these techniques by training the ResNet-50, Vision Transformer (ViT), and Swin Transformer V2 architectures, evaluating their effectiveness for this task. Our experiments revealed that the combination of ViT architecture and CutMix augmentation achieved the highest accuracy, with a performance of 0.9882.**

***Index Terms*—Spinal diseases, deep learning, data augmentation, CutMix, CutOut, MixUp.**

I. Introduction

The spinal column, or spine, plays a crucial role in human body movement, as well as protecting the spinal cord. Degenerative spinal pathologies, such as scoliosis and spondylolisthesis, significantly impact the quality of life, causing pain and movement restrictions [1], [2]. The National Health Survey, conducted by the Ministry of Health in partnership with The Brazilian Institute of Geography and Statistics (IBGE), revealed that 27 million adults in Brazil suffer from chronic spinal diseases, corresponding to 18.5% of the Brazilian adult population. Lumbar problems are the most common, with a higher prevalence in 21% of women compared to 15% of men. These numbers underscore the importance of more effective diagnostic tools to address these conditions that affect a significant portion of the population. [3]

Medical literature has focused on extracting these biomechanical parameters to quantify spinal diseases. However, approaches based on artificial intelligence (AI), such as [4], have emerged as promising alternatives, automating the diagnostic process. Recent studies employ deep convolutional neural networks and computer vision to diagnose these pathologies from radiographic images.

In this work, we propose an approach based on three deep learning models, ResNet-50 ViT and Swin Transformer, for classifying spinal X-ray images into (i) scoliosis, (ii) spondylolisthesis, and (iii) health. The impact of modern data augmentation strategies, such as CutMix, CutOut, and MixUp, are investigated in terms of improving the model's robustness and generalization while minimizing issues like overfitting.

Our results provide valuable insights into the capability of deep learning models to automatically discriminate between X-ray images of healthy patients and those obtained from patients with scoliosis and spondylolisthesis. These findings can be applied to support automated diagnostic solutions in clinical environments and application as automatic spinal curvature measurement [5], [6]. Moreover, computer vision techniques for detecting and classifying these conditions can lead to more accurate diagnoses, quicker interventions, and, consequently, a positive impact on the efficiency of medical resources.

Finally, our best result, with a mean accuracy of 0.9882, achieved by ViT trained with CutMix alongside standard data augmentation, surpasses the best result reported in the dataset original paper [7] of 0.9634 with DenseNet-201.

This paper is organized as follows. After this introductory section, we present and discuss some related works in Section II; In Section III, we describe the dataset and proposed methods. The results are presented and discussed in Section IV. Finally, we present our conclusions and perspectives in Section V.

II. Related Work

Deep learning models have shown promising results in automatically classifying spinal curvature, offering significant support in diagnosing scoliosis and spondylolisthesis. Several recent studies have focused on improving the accuracy of detecting these diseases through artificial intelligence techniques, replacing traditional methods prone to errors [8].

Fraiwan et al. [7] proposed an automatic classification system that eliminates manual measurements, utilizing deep learning models to distinguish between healthy, scoliosis, and spondylolisthesis. With a dataset containing 338 images, the system achieved an average accuracy of 96.73% in classifying

the three classes and over 98% in binary classifications. These results highlight the method's effectiveness as a diagnostic support tool, facilitating the early identification of pathologies and reducing the need for surgical interventions.

Zhang et al. [9] propose an application based on a deep learning model for classifying the severity of adolescent idiopathic scoliosis (AIS), identifying curve types, and monitoring disease progression. The model, named ScolioNets, was developed and validated using data from radiographs and back photographs of adolescents. Validation included comparing the model's results with the manual assessment by spine specialists. The results indicated that the model accurately classified AIS severity and distinguished different curve types.

Whang et al. (2021) [10] applied deep learning to spine X-rays to predict the progression of adolescent idiopathic scoliosis. An attentive capsule network was built to differentiate between progressive and non-progressive curve trajectories. A two-step transfer learning strategy was introduced to pretrain and fine-tuned the model. The model's performance was compared to CNNs and logistic regression methods. Camisa et al. (2024) [11] proposed an automated Computer-Aided Detection (CADe) system for spine lesions implemented using CT scans and the VGG-19 convolutional neural network. This neural network is capable of identifying healthy vertebrae and vertebrae with lesions. Finally, transfer learning techniques were used to achieve an accuracy of 93.43% and a recall of 92.99%.

Chen et al. [12] developed and validated a neural network approach to diagnose scoliosis from spinal radiographs. The proposed approach showed 95% accuracy in diagnosing scoliosis and can be used to support rapid diagnosis. Singh et al. [13] used a deep neural network trained on an X-ray dataset to classify vertebrae scans, proposing a model that extracts features from multiple layers and uses three fully connected layers for classification. The Singh et al. approach achieved 91.33% accuracy on the validation set and 92.22% on the test set. Lu et al. [14] applied transfer learning and fine-tuning to classify X-ray images of spinal diseases. The proposed approach is related to the Xception architecture with custom layers, which achieved a validation accuracy of 99.00%, a testing accuracy of 97.86%, and an F1-score of 97.86% in three-class classification (spondylolisthesis, scoliosis, and healthy).

Vergari et al. [15] present a classification algorithm to automatically detect scoliosis treatment in X-ray images that contain a brace, implant, or no treatment. The model uses a convolutional neural network (CNN) combined with discriminant analysis, enabling accurate image classification. A total of 796 X-rays of adolescents were analyzed, with the dataset augmented to 2,096 images through data augmentation techniques. Validation was performed using ten-fold stratified cross-validation, achieving an accuracy rate of 98.3%.

Although the studies above focused on classifying spinal X-ray images using deep learning models, such as CNNs, most did not explore the potential of different data augmentation strategies. This paper proposes an approach to improve classification generation by exploiting standard and advanced (including CutMix, CutOut, and MixUp) data augmentation strategies with different deep-learning models to address this gap.

III. Material and methods

A. Dataset

We utilized a dataset [16] [16][1] that comprises vertebral X-ray images to conduct experiments and evaluate training strategies. The dataset consists of X-ray images taken from 338 individuals, categorized into three classes: 71 individuals with healthy images, 79 diagnosed with spondylolisthesis, and 188 diagnosed with scoliosis. This dataset was collected at the King Abdullah University Hospital, Jordan University of Science and Technology, in Irbid, Jordan. The dataset is available with images in their original size and also with resized images for direct use in various deep-learning models. In this study, we employed the original-sized images.

The dataset presents several challenges, primarily due to its relatively small size. Additionally, variations in image quality and vertebral positioning in the X-rays increase the complexity of the automatic classification task. Another challenge involves accurately differentiating between the healthy, scoliosis, and spondylolisthesis classes, as the visual characteristics among the images can be minimal.

B. Architectures

For this study, we trained three deep learning architectures: ResNet-50 is a CNN, and Vision Transformer (ViT) and Swin Transformer are attention-based models.

Residual Network (ResNet) [17] is a CNN architecture recognized for introducing the concept of residual connections. The connections facilitate the efficient propagation of gradients through multiple layers, which minimizes the gradient vanishing problem.

The Swin Transformer [18] is a neural network architecture that combines the efficiency of convolutional networks with the ability to model long-range relationships typical of transformers. In short, this approach defines an image into segments that use a hierarchical sliding window structure to capture information at different scales.

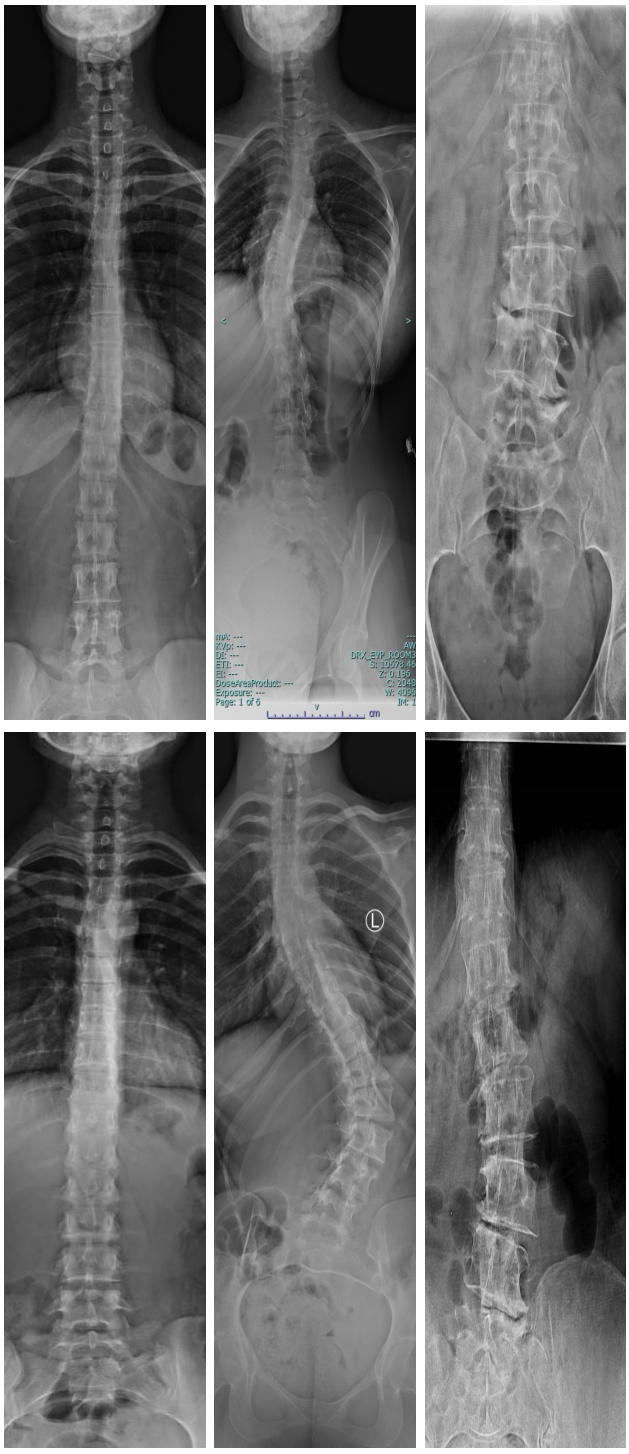
The Vision Transformer (ViT) [19] segments the input image into fixed blocks, then transforms the figures into vectors and processes them sequentially. In short, this architecture's characteristics enable capturing long-range dependencies in images.

C. Data augmentation

Data augmentation is a fundamental technique for dealing with small datasets, especially in applications where the limited number of images can compromise the generalization ability of deep learning models [20].

Data augmentation strategies enable the training process to learn more complex features from the same data and

[1] https://data.mendeley.com/datasets/xkt857dsxk/1



(a) healthy (b) Scoliosis (c) Spondylolisthesis

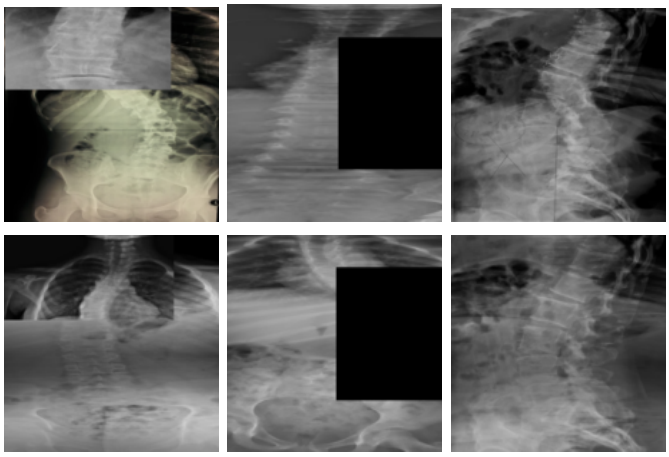
Fig. 1. Dataset images with respective classes: (a) healthy, (b) scoliosis, and (c) spondylolisthesis.

generate better models with generalization capability to avoid overfitting. In this work, we experiment with three advanced data augmentation strategies: CutOut, CutMix, and MixUp.

CutMix [21] is a data augmentation technique that merges information from two distinct images. During training, patches of one image are cut out and pasted onto another image within the same batch. The labels are then combined in proportion to the areas of the images. For instance, if 30% of one image is transferred to another, the resulting label will be composed of 70% from the original image and 30% from the inserted segment.

CutOut [22] is a simple data augmentation approach that involves randomly removing square regions from images during training. By cutting out parts of the visual information, the model is encouraged to focus on the most relevant features of the images rather than on specific details.

MixUp [23] is a data augmentation strategy that creates new training samples by combining two images and their respective labels. This technique generates an interpolation between two images and their labels. For example, a new image can be created using 60% of one image and 40% of another, with the labels adjusted proportionally.



(a) (b) (c)

Fig. 2. Samples of CutMix (a), CutOut (b), and MixUp (c) data augmentation applied over the input images after resizing to 224 $\times$ 224 pixels.

D. Experiment design

Three deep-learning classification models (ResNet-50, ViT, and Swin Transformer) were fine-tuned using the Adam optimizer with a learning rate of 0.0001 and batch size of 64. During training, the learning rate was adjusted using a cosine annealing strategy [24], which adjusts the learning rate over epochs, following a cosine-like function pattern. The training process was stopped early when the validation loss did not decrease for 21 consecutive epochs (patience) or if the training reached a maximum of 200 epochs.

All pre-trained models were obtained from the torchvision library, and all trainable parameters remained unfrozen during training, i.e., all network parameters were adjusted. The dataset

was first split into 70% for training and 30% for testing; after that, 25 % of the training set was split to become a validation set, leading to dataset partition of 50%, 20%, and 30% for training, validation, and testing, respectively. We choose to split 30% of the whole dataset for the test set based on the partition used in [7]. Moreover, all splits were performed in a stratified way.

Additionally, as we dealt with a very small dataset, each experiment was repeated five times, whereas, in each repetition, the dataset was randomly split, generating different dataset partitions. This strategy ensures a more reliable analysis of the models' performance over different dataset split scenarios. We then consider the mean and the standard deviation of the validation metrics computed over the test set. This decision was also taken to ensure a fair comparison with the work [7] that repeated each experiment X times with random dataset splitting.

Each model was trained with one of the three advanced data augmentation strategies (CutMix, CutOut, and MixUp) combined with a standard data augmentation strategy, totaling six experiments for each architecture. This strategy led to a set of eight experiments for each architecture.

For the training strategy without data augmentation, the images were only resized to 224 × 224 pixels, followed by normalization using the mean and standard deviation of the ImageNet dataset, as the models we used were pre-trained on it. The same transformations were applied for validation and testing.

For the training strategy with data augmentation, the training set images were subjected to random horizontal flipping, resizing to 224 × 224 pixels, random rotation of 15 degrees, subtle brightness, contrast, and saturation adjustments, followed by normalization using the mean and standard deviation.

E. Model evaluation

We analyzed the models trained on the test and validation sets considering the accuracy, as in Equations 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

F. Computational environment

The experiments were conducted on a PC with an i5 3.0 GHz processor, 32 GB of RAM, and an NVIDIA Titan Xp with 12 GB of memory. The development environment is based on Python 3.10, PyTorch 2.4.0, with CUDA 12.2. The libraries torchvision 0.19.0, scikit-learn 1.2.1, and Matplotlib 3.7.0 were also used.

IV. Results and discussion

Table I presents the test accuracies results, with mean and standard deviation obtained in the five randomized folds, for each model trained with both standard and advanced data augmentation strategies. In this table, it is possible to access the impacts of the advanced data augmentation strategies over each architecture when applied, with and without standard data augmentation procedures. The information present in Table I is also visually represented in Figure 3, in which it is possible to highlight the impact of the advanced data augmentation strategies on the trained models.

ResNet-50 achieved its highest accuracy of 0.9765 when trained without standard or advanced data augmentation techniques. The application of advanced augmentation methods reduced the model's performance. Specifically, CutMix slightly decreased accuracy to 0.9706, MixUp further reduced it to 0.9588, and CutOut resulted in the lowest accuracy of 0.9549. However, using standard data augmentation provided more consistent results, with MixUp improving its accuracy to 0.9725. This indicates that ResNet-50 benefits most from a straightforward approach and shows limited gains from advanced data augmentation techniques.

The ViT model demonstrated consistent performance, achieving an accuracy of 0.9765 both with and without standard data augmentation when no advanced augmentation techniques were used. The model benefitted from advanced augmentation strategies, except for MixUp, which led to the lowest accuracy of 0.9725 when used without standard data augmentation. When advanced techniques were combined with standard data augmentation, all showed improved results, with CutMix notably achieving the highest accuracy of 0.9882. This underscores ViT's effective use of advanced data augmentation methods to significantly enhance performance.

The Swin Transformer V2 delivered good results but exhibited greater accuracy variability than other models. The CutOut technique, when used without standard data augmentation, resulted in the best accuracy of 0.9804 among all experiments. However, applying MixUp with or without conventional data augmentation led to a decrease in accuracy to 0.9627 for both models. This indicates that the Swin Transformer V2 is more sensitive to certain data augmentation techniques.

The Swin Transformer V2 showed strong results but with higher accuracy variability. The CutOut technique, used without standard data augmentation, achieved the highest accuracy of 0.9804 among all Swin Transformer V2 models. In contrast, applying MixUp, regardless of whether standard data augmentation was used, resulted in a lower accuracy of 0.9627. This indicates the Swin Transformer V2's increased sensitivity to certain data augmentation methods.

Among the advanced data augmentation techniques, CutMix proved to be the most effective for the ViT b16 model, consistently enhancing accuracy across all configurations. Conversely, MixUp yielded the worst performance for the Swin Transformer V2 model. This suggests that advanced augmentation techniques are more advantageous for models with greater representational capacity, like the ViT b16, and for transformer-based architectures, like the Swin Transformer V2. In contrast, for models like the ResNet-50, standard data augmentation provides a satisfactory balance between data variation and training robustness.

TABLE I
MEAN AND STANDARD DEVIATION OF THE TEST ACCURACIES OBTAINED OVER THE FIVE RANDOMIZED FOLDS.

		Advanced Data Aug.			
Architecture	**Std. Data Aug.**	**None**	**CutMix**	**CutOut**	**MixUp**
ResNet-50	**No**	*0.9765* ± 0.0192	0.9706 ± 0.0139	0.9549 ± 0.0332	0.9588 ± 0.0259
	Yes	0.9706 ± 0.0248	0.9667 ± 0.0171	0.9588 ± 0.0209	*0.9725* ± 0.0190
ViT b 16	**No**	**0.9765** ± 0.0147	*0.9824* ± 0.0114	0.9804 ± 0.0062	0.9725 ± 0.0157
	Yes	**0.9765** ± 0.0118	***0.9882*** ± 0.0073	**0.9843** ± 0.0078	**0.9804** ± 0.0107
Swin T. V2 base	**No**	0.9647 ± 0.0253	0.9784 ± 0.0130	*0.9804* ± 0.0139	0.9627 ± 0.0300
	Yes	0.9725 ± 0.0227	*0.9784* ± 0.0144	0.9647 ± 0.0220	0.9627 ± 0.0280

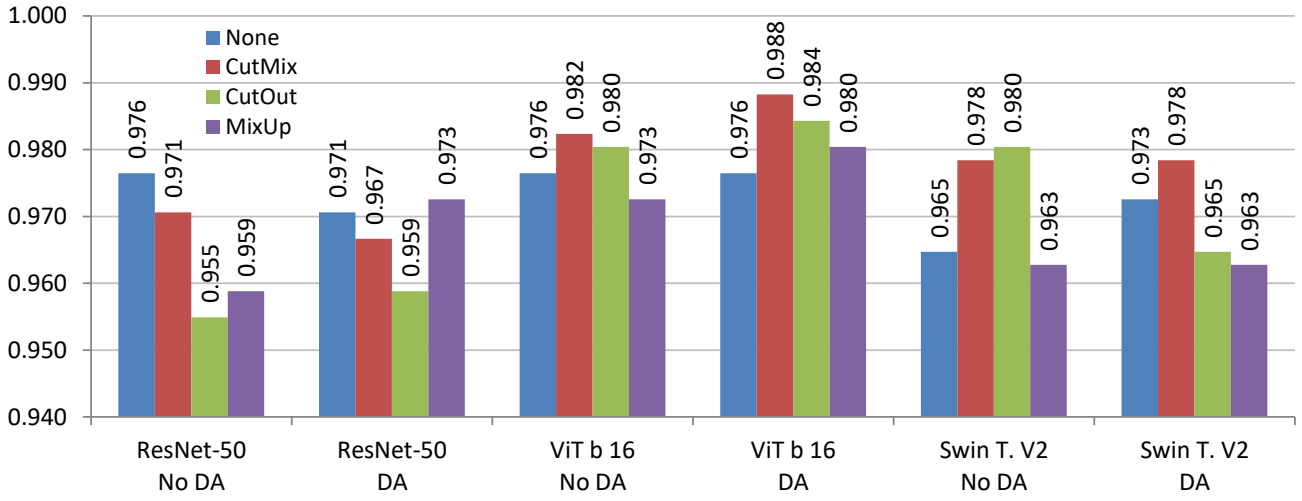


Fig. 3. Mean accuracy over the five randomized folds considering all training strategies.

V. CONCLUSION

In this work, we analyzed several deep learning models for classifying spinal X-ray images into three categories: healthy, scoliosis, and spondylolisthesis. We investigated the effects of advanced data augmentation techniques—namely CutMix, CutOut, and MixUp—both with and without standard data augmentation across three deep learning architectures: ResNet-50, ViT b16, and Swin Transformer V2. Our evaluation was conducted on a challenging dataset characterized by a limited number of images and highly imbalanced class distributions. This setup allowed us to assess how each data augmentation strategy influences model performance in this context.

The experiments conducted demonstrate that more modern architectures, such as ViT b16 and Swin Transformer V2, are more effective when leveraging advanced data augmentation techniques. ViT b16 stands out as the best option in terms of accuracy and consistency. Swin Transformer V2 is more sensitive to certain data augmentation methods, having the best results with CutOut alongside no standard data augmentation, followed by CutMix with standard data augmentation. On the other hand, ResNet-50, despite being a well-established architecture, performs best with simpler data augmentation techniques.

For future work, we propose leveraging Explainable Artificial Intelligence (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), to identify specific regions in spinal X-ray images that influence model classification decisions. This approach would enhance the interpretability of the model's outputs, offering clearer insights for diagnostic support. Grad-CAM, along with other XAI methods, could facilitate a deeper understanding of the neural network's behavior, allowing for more transparent validation of its decisions. Furthermore, integrating Grad-CAM with the fine-tuning of additional data augmentation strategies may enhance both the generalization and accuracy of deep learning models.

ACKNOWLEDGMENT

Omitted due to the double-blind review

REFERENCES

[1] American Association of Neurological Surgeons, "Scoliosis," https://www.aans.org/patients/conditions-treatments/scoliosis/, 2021, acessado: 22-ago-2024.

[2] American Academy of Orthopaedic Surgeons, "Spondylolysis and spondylolisthesis," https://orthoinfo.aaos.org/en/diseases--conditions/spondylolysis-and-spondylolisthesis, 2020, acessado: 22-ago-2024.

[3] Hospital de Cruzília, "Adultos com dor na coluna," https://www.hospitaldecruzilia.com.br/site/index.php/noticias/todas-as-noticias/586-adultos-dor-na-coluna, 2023, acessado: 22-ago-2024.

[4] I. Karpiel, A. Ziębiński, M. Kluszczyński, and D. Feige, "A survey of methods and technologies used for diagnosis of scoliosis," *Sensors*, vol. 21, no. 24, p. 8410, 2021.

[5] Y. Tu, N. Wang, F. Tong, and H. Chen, "Automatic measurement algorithm of scoliosis cobb angle based on deep learning," in *Journal of Physics: Conference Series*, vol. 1187, no. 4. IOP Publishing, 2019, p. 042100.

[6] M.-H. Horng, C.-P. Kuok, M.-J. Fu, C.-J. Lin, and Y.-N. Sun, "Cobb angle measurement of spine from x-ray images using convolutional neural network," *Computational and mathematical methods in medicine*, vol. 2019, no. 1, p. 6357171, 2019.

[7] M. Fraiwan, Z. Audat, L. Fraiwan, and T. Manasreh, "Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images," *Plos one*, vol. 17, no. 5, p. e0267851, 2022.

[8] B. H. Cho, D. Kaji, Z. B. Cheung, I. B. Ye, R. Tang, A. Ahn, O. Carrillo, J. T. Schwartz, A. A. Valliani, E. K. Oermann *et al.*, "Automated measurement of lumbar lordosis on radiographs using machine learning and computer vision," *Global spine journal*, vol. 10, no. 5, pp. 611–618, 2020.

[9] T. Zhang, C. Zhu, Y. Zhao, M. Zhao, Z. Wang, R. Song, N. Meng, A. Sial, A. Diwan, J. Liu *et al.*, "Deep learning model to classify and monitor idiopathic scoliosis in adolescents using a single smartphone photograph," *JAMA Network Open*, vol. 6, no. 8, pp. e2 330 617–e2 330 617, 2023.

[10] H. Wang, T. Zhang, K. M.-C. Cheung, and G. K.-H. Shea, "Application of deep learning upon spinal radiographs to predict progression in adolescent idiopathic scoliosis at first clinic visit," *EClinicalMedicine*, vol. 42, 2021.

[11] A. Camisa, G. Montanari, A. Testa, L. Falzetti, S. Avnet, N. Baldini, and G. Notarstefano, "Automated detection of spinal lesions from ct scans via deep transfer learning," *IEEE Access*, 2024.

[12] Q. Chen, R. Liao, M. Y. Shalaginov, and T. H. Zeng, "Scoliosis detection with convolutional neural networks," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 3785–3787.

[13] R. Singh, R. C. Joshi, A. Kumar, J. Singh, and M. K. Dutta, "Automated vertebrae diagnosis in spinal x-ray images using artificial intelligence," in *2023 4th International Conference for Emerging Technology (INCET)*. IEEE, 2023, pp. 1–6.

[14] Q. T. Lu and T. M. Nguyen, "An approach to classifying x-ray images of scoliosis and spondylolisthesis based on fine-tuned xception model." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.

[15] W. SKALLI, L. GAJNY, and C. VERGARI, "A convolutional neural network to detect scoliosis treatment in radiographs," 2020.

[16] M. Fraiwan, Z. Audat, and T. Manasreh, "A dataset of scoliosis, spondylolisthesis, and normal vertebrae x-ray images," 2022.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, vol. 137, p. 109347, 2023.

[21] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[22] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[23] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," *arXiv preprint arXiv:1905.08941*, 2019.

[24] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2022.