An Approach to Classifying X-Ray Images of Scoliosis and Spondylolisthesis Based on Fine-Tuned Xception Model

Quy Thanh Lu, Triet Minh Nguyen

Information Technology Department, FPT University, Can Tho, Viet Nam

Abstract—The vertebral column is a marvel of biological engineering and it considers a main part of the skeleton in vertebrate animals. In addition, it serves as the central axis of the human body comprising a series of interlocking vertebrae that provide structural support and flexibility. From basic works like bending and twisting to more complex actions such as walking and running, the spine's impact on human life is profound, underscoring its indispensable role in maintaining physical well-being and overall functionality. Moreover, in the hard-working schedule of people in modern life, a bunch of diseases impact on vertebral column such as spondylolisthesis and scoliosis. As a result, numerous researches were provided to take a hand in solving or avoiding these illnesses including machine learning. In this study, transfer learning and fine tuning were used for the classification of X-ray images on vertebrae sickness to avoid complex and wasted time in a medical examination process. The dataset for vertebrae illnesses X-ray images was collected at King Abdullah University Hospital and Jordan University of Science and Technology in Irbid, Jordan. It comprised 338 subjects including: 79 spondylolisthesis, 188 scoliosis, and 71 normal X-ray images. With the customized layers model in Xception that is used for image classification, we received surprisingly high results including validation accuracy, test accuracy, and F1 score in three-class classifications (i.e., spondylolisthesis, scoliosis, and normal) at 99.00%, 97.86%, and 97.86%, respectively. Additionally, two-class detection also received high accuracy values (i.e., 98.86% and 99.57%). Considering various high-performance metrics in the result indicates a robust ability to identify vertebrae diseases using X-ray images. The study found that machine learning significantly raises medical examinations compared to traditional methods, offering a myriad of benefits in terms of accuracy, efficiency, and diagnostic capabilities.

Keywords—Transfer learning; fine tuning; spondylolisthesis; scoliosis; classification; Xception

I. INTRODUCTION

In the busy world, many people deal with back problems like spondylolisthesis and scoliosis. These are issues with the spine that can make daily tasks such as working or studying harder. Spondylolisthesis happens when a vertebra in the spine moves out of place, causing pain in the lower back and sometimes putting pressure on nerves. Scoliosis is when the spine curves sideways in an unusual way. People with these conditions often have jobs that need a lot of physical effort, and sitting or standing for a long time can make things worse. To cope with these spine problems, many people use treatments that don't involve surgery. These include things like physical therapy, finding ways to manage pain, and making changes to the workspace to make it more comfortable. These steps help improve movement, reduce pain, and provide individuals continue contributing productively in their work.

Nevertheless, dealing with spondylolisthesis and scoliosis goes beyond just feeling uncomfortable. Moreover, it can also bring different levels of risk to people working even can be dangerous. If people do not take care of spondylolisthesis, it can lead to long-lasting pain, weak muscles, and even problems with nerves. Scoliosis, with its spine curvature, might cause breathing and heart issues and it affects overall health. In severe cases, surgery might be needed and make life more complicated. Despite these challenges, people should understand the management of their spinal conditions because it is crucial for motor nervous systems. In addition, it can help mitigate the risks associated with spondylolisthesis and scoliosis. This enables individuals to navigate their careers with resilience, adaptability, and a focus on their target.

Spinal diseases are increasing in modern times, especially scoliosis and spondylolisthesis. Despite this, symptoms of scoliosis with back pain are often overlooked by patients. In contrast, if people actively care about their health, we can easily identify the differences. Scoliosis and back pain seem to have specific characteristics in adult pain. For example, its location is often asymmetrical and associated with headaches. Furthermore, it is still unclear whether the intensity and duration of pain between adults with scoliosis and those without scoliosis experience back pain [1]. A lot of data collected in recent years around the world indicates the negative effects of scoliosis. To cite an example, the survey shows a dramatic rise in the average incidence of scoliosis diagnosis, climbing from 107 cases per 100,000 individuals in 2015 to 161 cases per 100,000 in 2022. Presently, approximately 1.2% of children and adolescents in Turkey are affected by scoliosis and the rate in women is 1.45 times higher than in men [2]. Besides, spondylolisthesis is a dangerous illness and it affects teenagers to elders. According to research degenerative lumbar spondylolisthesis affects 3% to 20% of globally and up to 30% of the elderly [3]. Additionally, research has shown that the illnesses are rare in those under 50 years old but they increase significantly with age affecting up to 15% of men and 50% of women aged 66–70 years [4]. This led to, people should pay more attention to their spine health to avoid future troubles.

X-ray images carry a pivotal role in the accurate diagnosis of spinal conditions, particularly in the classification of spondylolisthesis and scoliosis. X-ray images provide a comprehensive view of the spinal structure, enabling healthcare professionals to precisely identify and assess these conditions. However, X-ray imaging also has a bunch of limitations. Traditional methods rely heavily on manual interpretation, leading to the potential for human error and subjective variations in diagnosis. In addressing these challenges, machine learning emerges as a promising solution. By applying the power of artificial intelligence, machine learning algorithms can analyze vast datasets of x-ray images with high speed and accuracy. Nonetheless, it is crucial to acknowledge the limitations within the area of machine learning as well. The algorithms heavily depend on the quality and diversity of the training data, potentially leading to biased results. Furthermore, the interpretability of machine learning models in the medical field remains a challenge. Because of that, we need to improve machine learning models regularly to raise prediction and accuracy.

Artificial intelligence (AI) has come out as a trend in this day and age, particularly in the classification and segmentation of images. The ability of AI algorithms to categorize and organize huge datasets has transformed various industries [5], ranging from healthcare to finance. Machine learning techniques, such as deep learning and neural networks, have a main role in enhancing the accuracy and efficiency of classification tasks. These advancements have enabled AI systems to auto-recognize patterns, make predictions, and classify information with high precision. The integration of AI is also gaining prominence, addressing concerns about complex classification models such as illnesses on X-ray, MRI, and CT in health care [6]. In addition, if AI continues to develop its role will become unique in a new area where intelligent systems play an important role in decision-making processes across diverse domains. The trajectory of AI development in classification showcases its potential. This led to, we decided chose to develop the Xception model to gain high accuracy and solve more errors in X-ray image classification.

In this research, we use deep learning in the classification of images. In more detail, transfer learning was used to enhance performance in a novel task by utilizing knowledge acquired from previous learning experiences in similar tasks. By doing so, the model can capitalize on the generalized knowledge it has acquired, thereby improving its ability to tackle new challenges without starting from scratch [7]. Overall, transfer learning offers a powerful and efficient way to leverage previously acquired expertise, fostering improved performance and generalization across a range of tasks. In addition, fine-tuning is an important next step where the pre-trained model or its components are adjusted and optimized specifically for the new task [8]. This fine-tuning process ensures that the model adapts its learned features to the nuances of the target task, striking a balance between the general knowledge gained and the specifics of the current task. For this reason, we propose a method to use the Xception model in the Keras library in a Convolutional Neural Network (CNN) that uses transfer learning and fine-tuning to classify

images. Once trained, our model can classify new images or extract features for use in other applications such as object detection or image segmentation.

The contributions of this paper are as follows:

- Our research gains a high accuracy including validation accuracy, test accuracy, and F1 score in three classes' classifications in spondylolisthesis, scoliosis, and normal spine at 99.00%, 97.86%, and 97.86%, respectively. Moreover, pair-wise classification also achieves a high success up to 99.57%.

- Our study suggests a complete model that is used for vertebrae X-ray image classification including a dataset of scoliosis, spondylolisthesis, and normal vertebrae X-ray images. Thus, an expert can apply it in a simple way to help with the detection and classification of X-ray images.

- We find that Partition Explainer can be used effectively which is an algorithm that uses a hierarchical clustering of the data to recursively partition the input space.

- Our collected X-ray images of subjects with scoliosis and spondylolisthesis, as well as healthy ones, as determined by the specialists in the hospital This dataset is confirmed for the development of a model in deep learning including transfer learning and fine-tuning for the classification of vertebrae and can be applied to training and educating medical students, residents, and experts.

Our study comprises four main sections. Section II illustrates some of the related research that we used for references. Section III is the methodology, this section makes clear in detail all of the methods used in the article. Following that, Section IV will outline the experiments, detailing the methodology employed for conducting and assessing the accuracy of the deep learning model. Finally, we will provide a summary of our article and scrutinize the fundamental domains connected to the study in Section V.

II. RELATED WORK

An occupied working environment nowadays such as spending a lot of time at the working table or taking hours in the library to study. Based on several studies showing that, every year about 523 out of 100,000 teenagers develop scoliosis. This condition was twice as common in females compared to males based on the study population comprising 1782 teenagers from 10 to 18 years old [9]. Consequently, several researches on machine learning have been published for the segmentation and classification of X-ray images. For example, Peiji Chen et al. classified patient spine pictures using ResNet and Faster R-CNN. As a result, the combined use of ResNet convolutional neural network and Faster R-CNN has a stronger classification effect on scoliosis disorders than traditional machine learning approaches, as completely illustrated by the Area Under the Curve value of 90.87% [10]. Moreover, Joddat Fatima et al segmented the spinal column using Mask RCNN in conjunction with the YOLOv5 method for vertebral localization. The suggested method achieves 94.69% final average classification accuracy [11].

Machine learning plays a central role in classifying X-ray images for medical diagnosis. By leveraging algorithms, it can automatically identify patterns indicative of various conditions. This enhances diagnostic accuracy, expedites analysis, and contributes to more efficient and precise healthcare decision-making. Consequently, Shuman Han et al classified patients with moderate scoliosis with an accuracy of 77.9% and severe scoliosis with an accuracy of 93.6% using x-ray pictures of 204 patients with idiopathic scoliosis using the integrated area algorithm method of machine learning [12]. In addition, with a high accuracy of about 90.0%, Wahyu Caesarendra et al. suggest a deep learning architecture for the recognition of spine vertebrae from X-ray images [13]. This architecture automatically evaluates the Cobb angle and assesses for the presence of scoliosis and the severity of the curvature.

Especially in the analysis of X-ray pictures, Convolutional Neural Network (CNN) have completely changed deep learning for image classification. Their capacity is automatically extract hierarchical characteristics from pictures allows for the correct identification of patterns suggestive of different medical problems. CNN is required for improving X-ray image classification in medical diagnostics in terms of accuracy and precision. Furthermore, CNN is a common way to diagnose spondylolisthesis X-ray images in humans. For example, Fatih Varçın et al. used the MobileNet model in Convolutional Neural Network to classify spondylolisthesis or normal and achieved high results with a test accuracy reach of 99% [14]. Moreover, Deepika Saravagi et al. collected 229 X-ray images which include spondylolisthesis and the normal spine (i.e., 156 spondylolisthesis and 143 normal) which were optimized by applying the TFLite model optimization technique. As a result, the model reaches a high accuracy rate including the VGG16 model of 98% and InceptionV3 of 96% [15]. Additionally, Fatih Varçın et al. also AlexNet and GoogleLeNet models to classify the data set consisting of 272 X-ray images. According to experimental results, GoogleLeNet performs marginally better than AlexNet, which has an accuracy of 91.67%, with a 93.87% accuracy rate [16].

Processing medical images in X-ray images has witnessed significant promotions through the utilization of transfer learning and fine-tuning techniques. Leveraging pre-trained models allows the transfer of knowledge from general domains to medical imaging while fine-tuning tailors the model for specific diagnostic tasks. This approach enhances the efficiency and effectiveness of X-ray image analysis in medical applications. For instance, Mohammad Fraiwan et al. used transfer learning in the DensNet-201 model and reached a mean accuracy and maximum accuracy for spine illness classification were 96.73% and 98.02%, respectively [17]. Furthermore, Using the VGG16 model for feature extraction and CapsNet for disease identification, Deepika Saravagi's experimental results show 98% accuracy [18]. The dataset contains 466 X-ray radiographs, with 186 images showing a spine with spondylolisthesis and 280 images showing a normal spine.

Deep learning models could help handle the growing amount of medical imaging data and offer an early analysis of pictures collected in basic care. When it comes to scoliosis identification, deep learning algorithms provide a faster and more effective solution than manual X-ray investigation. Arslan Amin et al. used a pre-trained EfficientNet model to achieve an accuracy of 86 % on the detection and classification of scoliosis from X-ray images [19]. Besides, Ariana Alejandra Andrews Interiano et al. take a database of medical images from Honduran to transfer learning and fine-tuning in InceptionResNet, MobileNet, and EfficientNet. Hence, their experiment finds a high average accuracy of 98.01% [20]. Furthermore, Dalwinder Singh et al. applied CNN to classify MRI lumbar spine images and used differential spider monkey optimization (SMO) to get the highest classification accuracy of 96% [22]. In conclusion, a bunch of different research has been published in recent times to propose the accuracy in segmentation and classification in medical and help patients avoid a lot of time and money for a long procedure in treatment.

III. METHODOLOGY

A. The Research Implementation Procedure

This study proposes a method including 12 steps shown in Fig. 1. The roles of the steps are shown as follows:
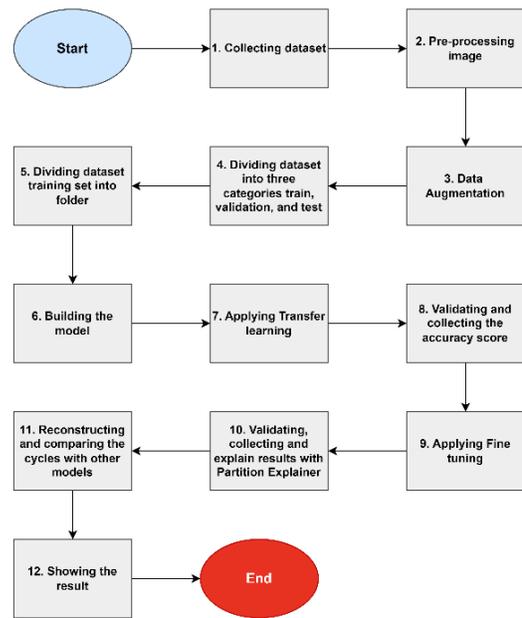

Fig. 1. The implementing procedure flowchart.

1) Collecting dataset: The dataset about vertebrae illnesses is collected at King Abdullah University Hospital and Jordan University of Science and Technology in Irbid, Jordan. The collection contains X-ray images of two types of spine illness that is spondylolisthesis, and scoliosis. Besides, one class for normal images is provided. This collection provides a valuable resource for medical research.

2) Pre-processing image: Standardized input conditions were fixed for CNN models through the use of resizing and normalization. As a result, the outcomes of the results grow.

3) Data augmentation: This step is a technique of artificially increasing the dataset by creating modified copies

of a dataset using existing data to apply functions such as rotate, flip, and brightness contrast.

4) Dividing the dataset into three categories train validation and test: The entire X-ray images dataset includes 3500 subjects after increasing in data augmentation by 338 default subjects with random selection used in the phases of training, validation, and testing. An 8-1-1 scale is used to randomly choose the datasets, dividing them into eight halves for training, 1 for validation, and 1 for testing. This ensures a balanced distribution, which is necessary for reliable model creation and assessment.

5) Dividing the training set into folders: spondylolisthesis, scoliosis, and normal spine are divided into many different folders. The first folder is 3-fold including spondylolisthesis, scoliosis, and normal spine. Our goal is to compare the largest folder by displaying the training data in a more precise manner. As a result, the other folder has 2-fold classifications: scoliosis-normal and spondylolisthesis-normal.

6) Building the model: To do experiments, we used transfer learning to a pre-trained model and rebuilt the model based on the CNN architecture prototype. Subsequently, fine-tuning is the process of modifying the weights of the pre-trained model on the particular data of the target job. Consequently, the Xception model produces an outstanding outcome for our training test.

7) Applying transfer learning: In transfer learning, a large dataset was used for leveraging a pre-trained model. This dataset may contain a large amount of labeled data. By using knowledge gained from the source task, transfer learning enhances the performance of the model on the target task, particularly when data for the latter is limited.

8) Validating and collecting the accuracy score: We summarized the training accuracy obtained from the predictions made by the model to evaluate its accuracy after it had finished training. Next, we used the initially divided testing set to assess whether the test was correct.

9) Applying Fine tuning: Fine-tuning was applied to the act of modifying the parameters of a pre-trained neural network and the hyperparameters of a model to improve its performance, often in the last layers. This enables the model to draw on elements learned in a broader context while customizing its knowledge to the specifics of the target task.

10) Validating, collecting and explain results with Partition Explainer: After collecting all the metrics such as validation accuracy, test accuracy, and F1 score. After that, a partition explainer in SHAP was used for a specific algorithm for explaining the output of machine learning models. SHAP is a unified approach to explaining the output of any machine learning model, and it is based on Shapley values from cooperative game theory.

11) Reconstructing and comparing the cycles with other models: After the first phase, we rework and compare it with another model including EfficientNetB3, VGG19, ResNet101, and DenseNet169 to create the final result.

12) Showing the result: Following a comparison, the data will be displayed in the form of tables and graphs to allow for relevant comparisons.

B. Pre-processing Image

In the area of image processing, the pre-processing stage plays a central role in enhancing the efficiency and effectiveness of subsequent tasks, such as machine learning model training. Two fundamental operations within this pre-processing pipeline are image resizing (1) and normalization (2). Image resizing involves transforming the dimensions of an image, commonly to a standardized size, to facilitate uniformity and computational feasibility.

The resizing operation is typically represented by the formula:

$$Resized\ Image = Resize\big(Original\ Image, (224,224)\big)\# \tag{1}$$

In Formula , the original image goes through a transformation to conform to a predefined resolution of 224×224 This standard size is often used to ensure consistency across the dataset and compatibility with neural network architectures commonly used in computer vision tasks.

Following resizing, the next critical step is normalization, see Formula (2), a process focused on normalizing the pixel values of the image. Normalization is carried out to ensure that the input data falls within a specific range, which aids in stabilizing the learning process during model training. The normalization operation can be mathematically expressed as:

$$Normalized\ Image = \frac{Resized\ Image - \min(Resized\ Image)}{\max(Resized\ Image) - \min(Resized\ Image)}\# \tag{2}$$

Here Formula (2), the pixel values of the resized image are transformed to a range between 0 and 1 by subtracting the minimum pixel value and dividing by the range between the maximum and minimum pixel values. This normalization to the [0, 1] range is crucial for mitigating issues related to varying scales and ensuring that the model receives consistent input across diverse images.

In summary, the connection of resizing and normalization in image pre-processing not only standardizes the size of input images but also establishes a common pixel value scale.

C. Data Augmentation

Augmenting data is a critical step in improving the robustness and generalization capability of machine learning models, notably in picture classification. One widely used strategy involves applying several changes to the original images, resulting in a diverse set of training samples for the model to learn from.

The first step in Formula (3) is the transpose operation involves swapping the rows and columns of the image matrix. Mathematically, if we have an image represented by a matrix I of dimensions $m \times n$, where m is the number of rows and n is the number of columns, the transpose has denoted as I_{trans} (3)

results in a new matrix with dimensions $n \times m$, it can be expressed as:

$$I_{trans} = I^T \#\qquad(3)$$

The next step in Formula (4), shift scale rotate is used for translations, scaling, and rotations to the image. The rotated image I_{rot} is obtained by applying a rotation matrix $R(\theta)$ (4) to the original image matrix I:

$$I_{rot} = R(\theta) \cdot I \#\qquad(4)$$

Here, the rotation limit is set to 45 degrees ($\theta \leq 45$) with a probability p = 0.45 for each image. This ensures a controlled augmentation process that is both effective and computationally efficient.

The third step , horizontal flip (5) and vertical flip (6) operations involve mirroring the image horizontally and vertically, respectively. By means of mathematics, the horizontal flip $I_{horizontal\_flip}$ (5) is achieved by reversing the order of columns in the original image matrix I, and the vertical flip $I_{\text{vertical\_flip}}$ (6) is achieved by reversing the order of rows:

$$I_{horizontal\_flip} = flip(I, axis = 1)\#\qquad(5)$$

$$I_{\text{vertical\_flip}} = flip(I, axis = 0)\#\qquad(6)$$

Both operations are applied with a probability of $p = 0.5$ to introduce variability in the orientation of the training samples.

The final step (7) is one operation of random brightness contrast and it can be expressed as a single formula, where the brightness and contrast adjustments are applied to each pixel in the image:

$$I_{bc} = I + r_{brightbess} \times I \times r_{contrast}\#\qquad(7)$$

In this formula, I_{bc} (7) represents the image after the combined brightness and contrast adjustments, I is the original image matrix. Moreover, $r\_brightbess$ (7) is a randomly sampled value for brightness adjustment, and $r\_contrast$ (7) is a randomly sampled value for contrast adjustment.

These adjustments are executed with a probability of $p = 0.2$ to ensure controlled variability without excessively distorting the image characteristics. In summary, these augmentation techniques collectively contribute to a more diverse and robust dataset, fostering improved performance and generalization of machine learning models in image classification tasks.

D. Transfer Learning and Fine Tuning of Xception

Transfer learning and fine-tuning are powerful techniques in the area of CNN [22], [23], allowing the utilization of pre-trained models to enhance the performance of a specific task. One noteworthy architecture for such applications is the Xception model, which stands out for its depth and efficient use of parameters. Unlike traditional CNN, Xception uses an extreme version of the inception module, known as the depthwise separable convolution. This technique separates the spatial and channel-wise operations, enabling the model to capture both local and global features effectively.
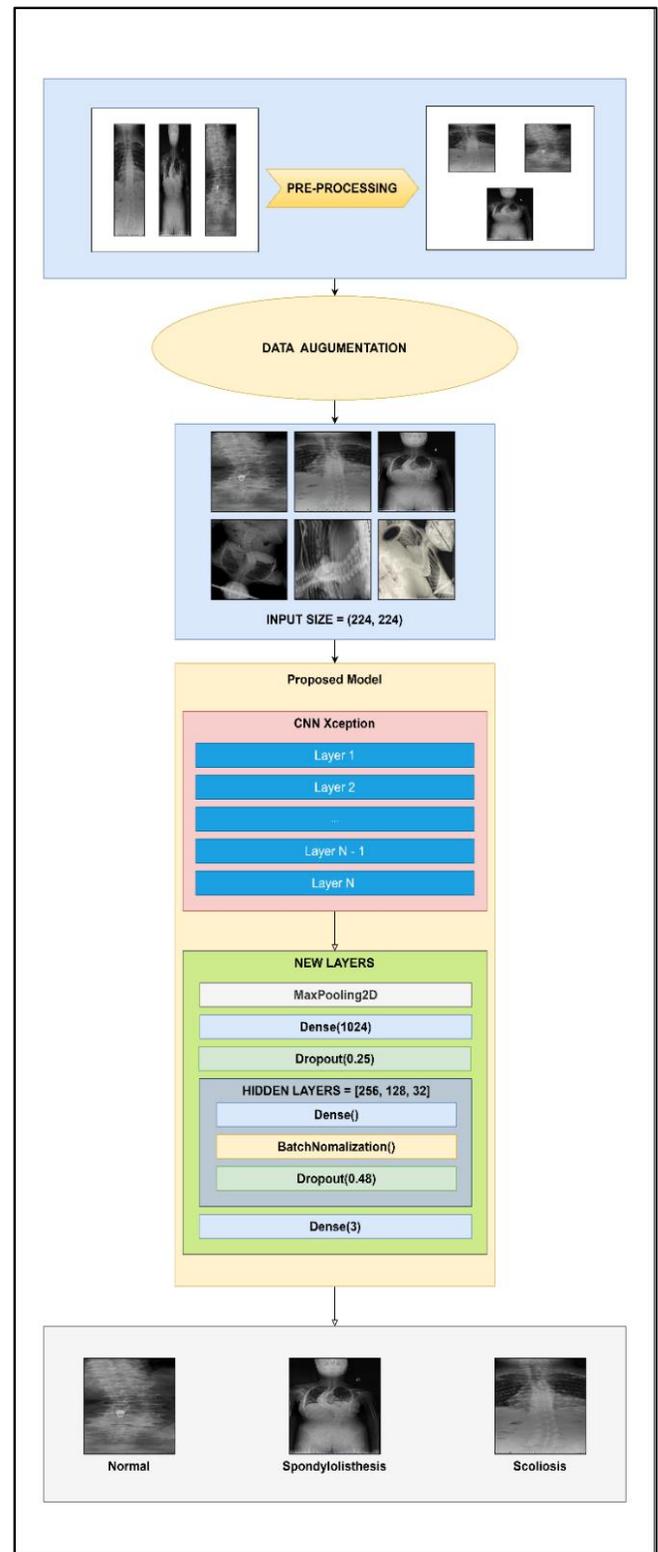

Fig. 2. Procedure of transfer learning and fine-tuning in CNN Xception model and custom layers.

The Xception model, introduced by François Chollet in 2017, is an extension of the Inception architecture. Its key innovation lies in replacing standard convolutions with depthwise separable convolutions, resulting in a more efficient

and parameterized model. This architectural shift reduces the risk of overfitting, enhances feature representation, and facilitates faster training convergence. Each depthwise separable convolutional block in Xception consists of a depthwise convolution followed by a pointwise convolution, providing a powerful yet lightweight alternative to conventional convolutional layers.

When it comes to transfer learning and fine-tuning in CNNs, the Xception model proves particularly advantageous. Leveraging the pre-trained weights from a large dataset, such as ImageNet, Xception can be employed as a feature extractor for a diverse range of computer vision tasks. Our study adds more external layers to increase accuracy in Fig. 2 and this process not only saves computational resources but also leverages the rich hierarchical features learned by Xception, enhancing the model's ability to generalize across various visual patterns.

In essence, the seamless integration of the Xception model into CNN architectures and the addition layer described in Fig. 2 for transfer learning and fine-tuning extends the paradigm of leveraging pre-trained models, unlocking the potential for enhanced performance and efficiency in a myriad of computer vision applications.

E. Explain Results with Parition Explainer

In the study, Partition Explainer a method within SHAP (Shapley additive explanations) was chosen as a visual explanation. It serves as a necessary tool in explaining the contributions of individual features in an image-based model. This process is particularly useful for understanding the importance of different aspects within an image and gaining insights into model decision-making. At its core, the Partition Explainer leverages Shapley values, a concept rooted in cooperative game theory, to fairly distribute the model's output among its input features.

In more detail, the Partition Explainer operates by considering all possible subsets of features and calculates the average Shapley value (8) for each feature across these subsets. This careful approach ensures a comprehensive evaluation of the impact of each feature, accounting for their interactions and dependencies. Mathematically, the Shapley value for a feature (φ_i) (8) in a cooperative game is expressed as follows:

$$\varphi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|\,!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad \# \quad (8)$$

In this Formula (8), N represents the set of all features, S denotes a subset of features excluding 'i', and f(S) signifies the model's output when considering the subset of features 'S'. The Shapley value quantifies the marginal contribution of feature 'i' by averaging across all possible combinations, providing a fair and consistent measure of its impact on the model's output.

In the context of the Partition explainer, this Shapley value calculation is extended to various feature subsets, enabling an expression understanding of how each feature influences the model's predictions.

By using Partition explainer in the final result of Fig. 3, our results gain insights into model behavior, fostering trust and facilitating informed decision-making in the area of machine learning.
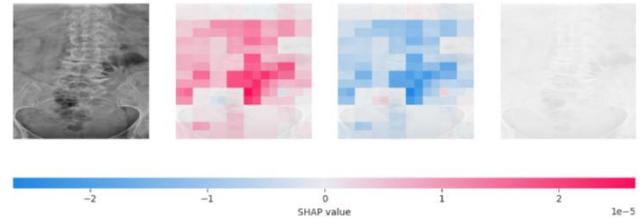

Fig. 3. The final result of classification spondylolisthesis after applying a partition explainer.

IV. EXPERIMENTS

A. Dataset and Peformance Metrics

For this analysis, a single dataset Fig. 4 was used for training, validation, and testing. A total of 338 pictures, comprising 79 spondylolisthesis, 188 scoliosis, and 71 normal, make up the full X-ray images dataset that was obtained and enhanced by King Abdullah University Hospital and Jordan University of Science and Technology in Irbid, Jordan. The dataset increased to 3500 pictures after data augmentation and it was divided into 8 for training, 1 for validation, and 1 for testing.

Fig. 4. Dataset about the vertebrae X-ray images.

Additionally, five measures were used to evaluate the model performance: the F1 score, test accuracy, recall, precision, and validation accuracy all have a significant impact on how well a trained model performs and its capacity for generalization.

The F1 score in Formula (9) is the harmonic mean of precision and recall. It balances the trade-off between precision and recall, providing a single value that takes both false positives and false negatives into account. The F1 score is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \# \quad (9)$$

Test accuracy as in Formula (10) measures the proportion of correctly predicted instances over the total number of instances in the test set. It is a common metric for overall classification performance, providing insights into its real-world applicability. This metric is calculated by:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \# \quad (10)$$

Validation accuracy in Formula (10) is similar to test accuracy, it measures the proportion of correctly predicted instances over the total number of instances in the validation

set. It is used during the training process to monitor the model's performance on a separate dataset not used for training.

Precision in Formula (11) talks about the accuracy of positive predictions made by the model, emphasizing minimizing false positives. The precision formula is given by:

$$Precision = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives} \#$$ (11)

Recall in equation (12), a metric crucial in scenarios where identifying true positives is paramount, is defined as:

$$Recall = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives} \#$$ (12)

B. Scenario 1: The Results of Classifying X-Ray Images into Two Classes: Scoliosis and Normal Spine

Through customization and training, the scenario aimed to assess how well the pre-trained models performed in correctly diagnosing the X-ray image condition. Furthermore, by using these statistics, we may more easily and intuitively compare the vertebral X-ray images in three classes: normal spine, scoliosis, and spondylolisthesis.

Table I show the performance evaluation metrics for classifying in two classes. The ResNet101 achieved the highest accuracy value in transfer learning over the two statistical measures with a validation accuracy of 99.14%. Test accuracy, precision, and F1 score all reached 98.29%.

On the other hand, in Table II TABLE II. shows our model performed the best after fine-tuning with a validation accuracy of 98.86%. Test accuracy, precision, and F1 score all reached 99.14%. This led to, underscore the effectiveness of customizing the model to the nuances of the target task through fine-tuning. This suggests that while transfer learning provides a strong foundation, fine-tuning allows for a more tailored approach, particularly when dealing with domain-specific nuances.

A sample training and validation progress curve with the loss and accuracy values of our model during fine-tuning is displayed in Fig. 5 and Fig. 6. The graphic shows suitable training and validation sets along with consistent learning behavior. Thus, it shows how our work's fine-tuning accuracy has increased.

TABLE I.   THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO TWO CLASSES: NORMAL SPINE AND SCOLIOSIS IN TRANSFER LEARNING, FOR EACH DEEP LEARNING MODEL

Model	Transfer learning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	99.00%	97.71%	97.74%	97.71%	97.72%
DenseNet169	88.29%	87.43%	87.24%	87.43%	87.31%
VGG19	94.29%	80.14%	90.86%	90.86%	90.86%
ResNet101	99.14%	98.29%	98.29%	98.29%	98.29%
Our Model	**87.00%**	**83.86%**	**84.33%**	**83.86%**	**84.05%**

TABLE II.   THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO TWO CLASSES: NORMAL SPINE AND SCOLIOSIS IN FINE TUNING, FOR EACH DEEP LEARNING MODEL

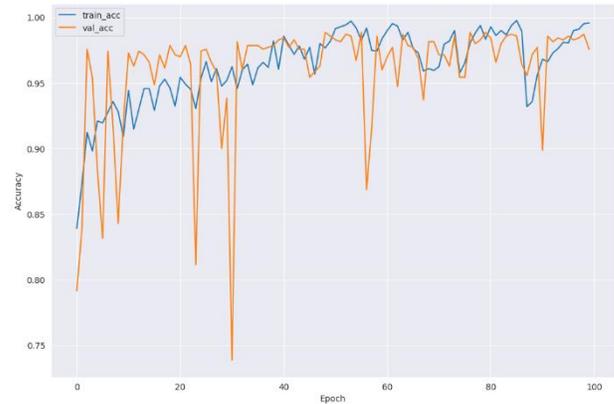| Model	Fine tuning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	97.71%	97.43%	97.42%	97.43%	97.42%
DenseNet169	96.00%	94.00%	93.94%	94.00%	93.95%
VGG19	73.71%	73.57%	54.13%	73.57%	62.37%
ResNet101	79.57%	78.57%	77.09%	78.57%	75.90%
Our Model	**98.86%**	**99.14%**	**99.14%**	**99.14%**	**99.14%**

Fig. 5.   Training accuracy and validation accuracy in fine-tuning in two classes normal and scoliosis of our model.
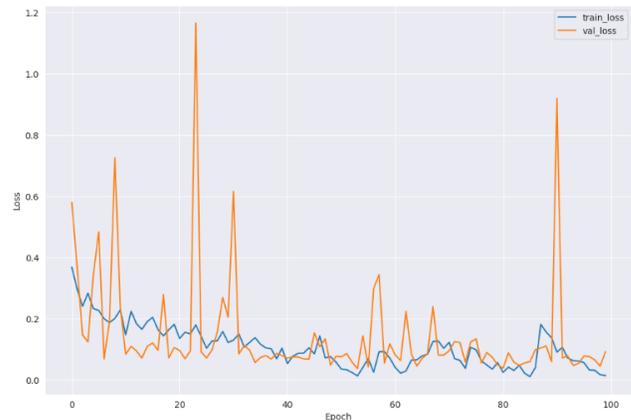

Fig. 6.   Training loss in and validation loss fine-tuning in two classes normal and scoliosis of our model.

Fig. 7 indicates the confusion matrix of two-class (i.e., scoliosis and normal spine) in 700 pictures.

C. Scenario 2: The Results of Classifying X-Ray Images Into Two Classes: Spondylolisthesis and Normal Spine

Table III indicates the highest result of EfficientNetB3 in transfer learning (i.e., 100%) and other models also achieved a high result (i.e., > 95%). Moreover, Table IV shows a reduction in the results of EfficientNetB3, VGG19, and ResNet101 but our model has signification growth (i.e., 99.43%).
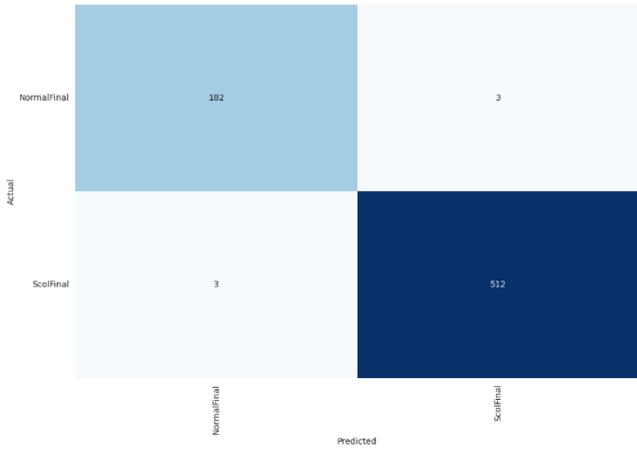
Fig. 7. Confusion matrix in fine-tuning in two classes normal and scoliosis of our model.

TABLE III. THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO TWO CLASSES: NORMAL SPINE AND SPONDYLOLISTHESIS IN TRANSFER LEARNING, FOR EACH DEEP LEARNING MODEL

Model	Transfer learning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	99.86%	100.00%	100.00%	100.00%	100.00%
DenseNet169	97.57%	97.71%	97.72%	97.71%	97.72%
VGG19	99.71%	99.57%	99.57%	99.57%	99.57%
ResNet101	99.86%	99.86%	99.86%	99.86%	99.86%
Our Model	**96.43%**	**95.57%**	**95.59%**	**95.57%**	**95.57%**

TABLE IV. THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO TWO CLASSES: NORMAL SPINE AND SPONDYLOLISTHESIS IN FINE TUNING, FOR EACH DEEP LEARNING MODEL

Model	Fine tuning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	99.86%	99.14%	99.14%	99.14%	99.14%
DenseNet169	99.14%	98.86%	98.87%	98.86%	98.86%
VGG19	93.29%	91.43%	91.50%	91.43%	91.44%
ResNet101	97.14%	97.00%	97.03%	97.00%	97.00%
Our Model	**99.57%**	**99.43%**	**99.43%**	**99.43%**	**99.43%**

Fig. 8 and Fig. 9 in this experiment explain training accuracy and training loss in our model for two classes of normal and spondylolisthesis which low test loss (i.e., ~0%).

The outcome confusion matrix is finally displayed in Fig. 10, demonstrating the excellent performance of our model.

D. Scenario 3: The Results of Classifying X-Ray Images into Three Classes: Spondylolisthesis, Scoliosis, and Normal Spine

Table V and Table VI illustrate in transfer learning the ResNet101 reaches the highest accuracy value over the three statistical measures with a validation accuracy of 99.00%, test accuracy of 97.71%, and F1 score of 97.72%. However, our model performed achieved the lowest rank in transfer learning with a validation accuracy of 82.00%, test accuracy of 80.71%,

and F1 score of 79.97%. The final result is only improved in fine-tuning after our research added more layers and that proves our achievements exactly when our model gets a validation accuracy of 99.00%, test accuracy of 97.86%, and F1 score of 97.86%.
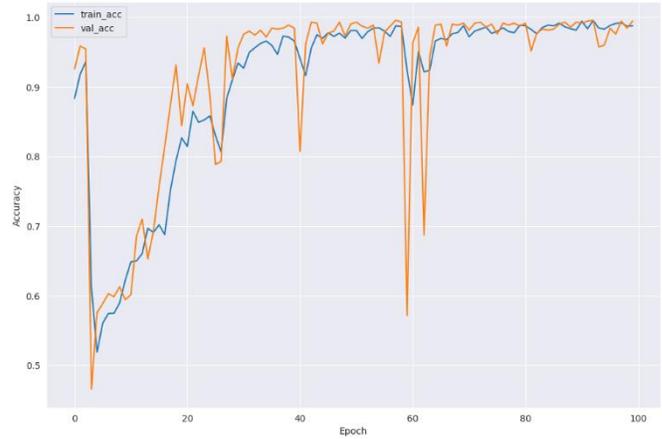

Fig. 8. Training accuracy and validation accuracy in fine-tuning in two classes normal and spondylolisthesis of our model.
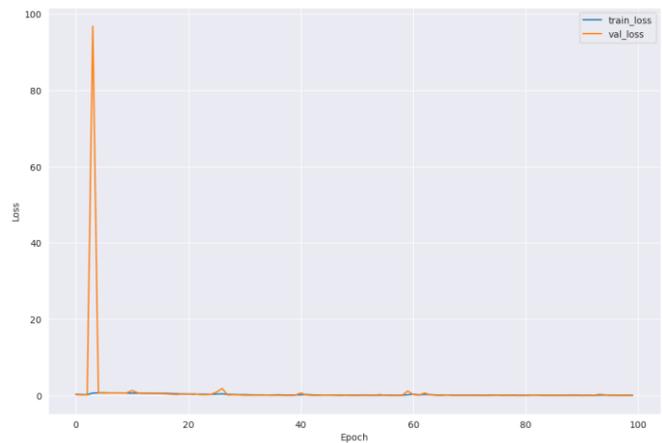

Fig. 9. Training loss in and validation loss fine-tuning in two classes normal and spondylolisthesis of our model.
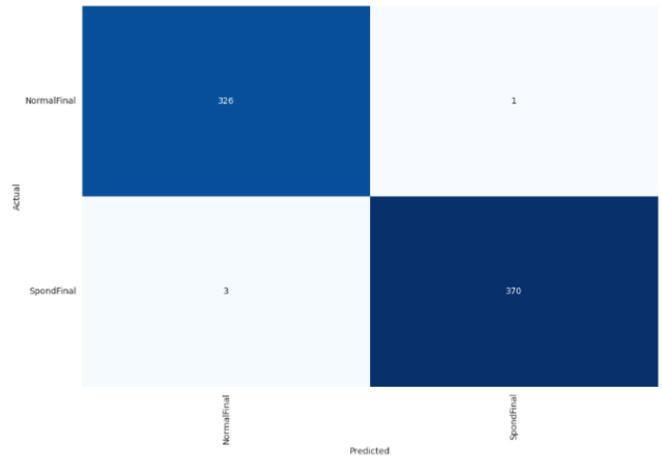

Fig. 10. Confusion matrix in fine-tuning in two classes normal and spondylolisthesis of our model.

TABLE V. THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO THREE CLASSES: NORMAL SPINE, SCOLIOSIS, AND SPONDYLOLISTHESIS IN TRANSFER LEARNING, FOR EACH DEEP LEARNING MODEL

Model	Transfer learning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	98.86%	97.71%	97.77%	97.71%	97.73%
DenseNet169	88.43%	86.43%	87.27%	86.43%	86.67%
VGG19	91.86%	64.43%	89.40%	89.00%	89.03%
ResNet101	99.00%	97.71%	97.74%	97.71%	97.72%
Our Model	**82.00%**	**80.71%**	**80.45%**	**80.71%**	**79.97%**

TABLE VI. THE ACCURACY OF CLASSIFYING X-RAY IMAGES INTO THREE CLASSES: NORMAL SPINE, SCOLIOSIS, AND SPONDYLOLISTHESIS IN FINE TUNING, FOR EACH DEEP LEARNING MODEL

Model	Fine tuning				
	Val acc	Test acc	Precision	Recall	F1
EfficientNetB3	96.86%	96.71%	96.75%	96.71%	96.72%
DenseNet169	95.71%	93.57%	93.65%	93.57%	93.60%
VGG19	69.57%	26.00%	53.17%	67.00%	59.10%
ResNet101	84.29%	82.57%	82.74%	82.57%	82.14%
Our Model	**99.00%**	**97.86%**	**97.88%**	**97.86%**	**97.86%**

The training and validation progress curves for a scenario run of the best-performing model are displayed in Fig. 11 and Fig. 12. A model's performance on the training data is measured by training accuracy, which indicates how well the model can learn from the given instances.

The model is guided to reduce mistakes during training by measuring the difference between anticipated and actual values in the training set, which is known as training loss. Validation loss is a crucial metric for assessing the generalization performance of the model, replicating this procedure on an independent dataset.

Fig. 12 shows the sample confusion matrix for three classes of classification. This important step makes it possible for us to see a more intuitive comparison of the results achieved. Fig. 13 illustrates the final result displayed in the SHAP value of Partition Explainer in Fig. 14 which is an excellent way to present visually and provides an overall view for experts and medical teams.

E. Comparison with others State-of-the-art Methods

To examine the accuracy of the proposed model that our article has just given out in the previous section, we compare the accuracy score of the proposed model with other CNN architectures in Table VII, which are EfficientNetB3, DenseNet169, VGG19, and ResNet101.

Our comparison serves as a standardized benchmark, allowing researchers to evaluate the performance of new approaches, identify strengths and weaknesses, and push the boundaries of what is achievable. This process fosters healthy competition, driving innovation and motivating the community to build upon successful methodologies. Assessing

generalization across diverse datasets, understanding resource utilization, and uncovering limitations are key outcomes of such comparisons. Moreover, it ensures reproducibility, aligns research with community standards and guides future endeavors toward addressing challenges and improving the overall state of the art in deep learning.
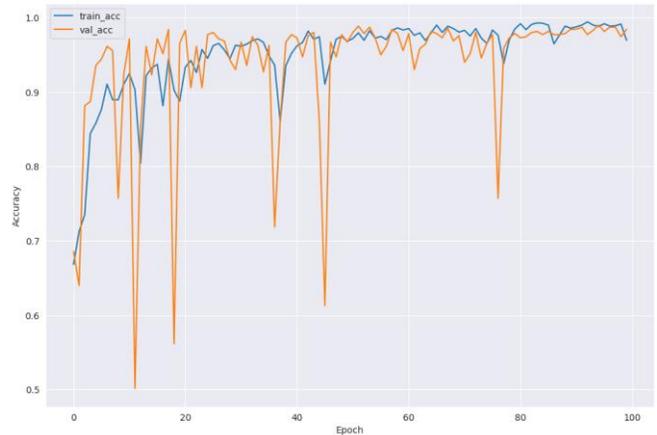

Fig. 11. Training accuracy and validation accuracy in fine-tuning in three classes of our model.
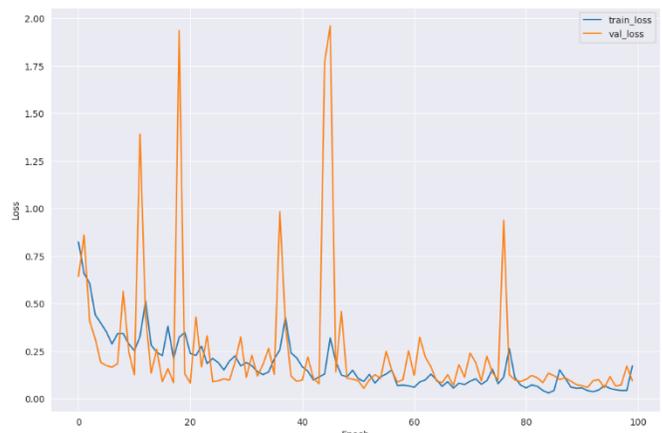

Fig. 12. Training loss in and validation loss fine-tuning in three classes of our model.
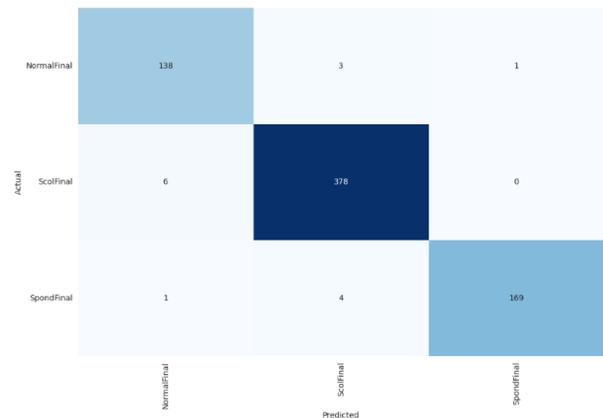

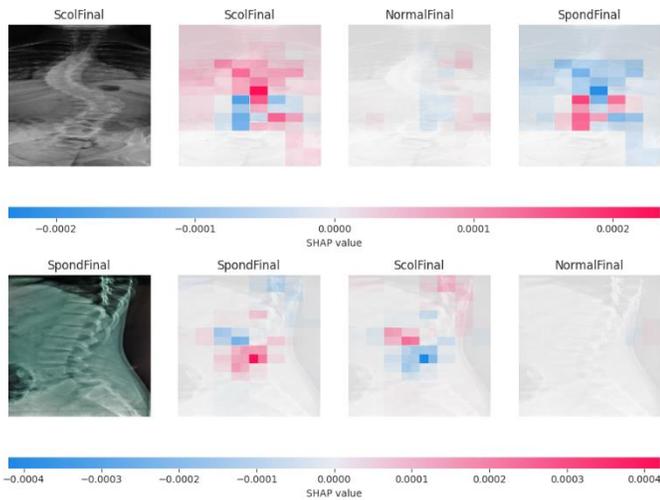Fig. 13. Confusion matrix in fine-tuning in three classes of our model.

Fig. 14. The final result with Partition Explainer.

TABLE VII. COMPARISON WITH OTHERS STATE-OF-THE-ART METHODS

Ref.	Proposed	Accuracy
Peiji Chen et al.	ResNet and Faster R-CNN	90.87%
Joddat Fatima et al.	Mask RCNN and YOLOv5	94.69%
Shuman Han et al.	ROC and Cobb	77.9% ~ 93.6%
Wahyu Caesarendra et al.	AutoSpine-Net	90.00%
Fatih Varçın et al.	MobileNet	99.99%
Deepika Saravagi et al.	InceptionV3 and VGG16	96.00% ~ 98.00%
Fatih Varçın et al.	AlexNet and GoogleLeNet	91.67%
Mohammad Fraiwan et al.	DensNet-201	96.73% ~ 98.02%
Deepika Saravagi et al.	VGG16	98.00%
Arslan Amin et al.	EfficientNet	86.00%
Ariana Alejandra Andrews Interiano et al.	InceptionResNet, MobileNet, and EfficientNet	98.01%
Dalwinder Singh et al.	SMO	96.00%
Proposed model		**97.86%**

V. CONCLUSION

Our newly developed model showcases commendable performance in classifying vertebrae X-ray images, specifically distinguishing between normal spines, scoliosis, and spondylolisthesis for critical medical applications. The model exhibits a remarkable validation accuracy of 99.00%, a robust test accuracy of 97.86%, and an F1 score of 97.86%, underscoring its efficacy in accurately identifying and categorizing spinal conditions. The success of our model can be attributed to strategic modifications, including the incorporation of dense and dropout layers into the Xception model, coupled with fine-tuning various settings, resulting in a substantial improvement in overall accuracy.

Transfer learning played a pivotal role in our approach, leveraging the pre-trained Xception model as a foundation. This technique involves utilizing knowledge gained from a task-specific source domain, in this case, the general image recognition capabilities of the Xception model, and applying it to our specific task of vertebrae classification. Fine-tuning further refined the model's performance by adjusting its parameters to align with the intricacies of our dataset. This process enhances the model's ability to discern subtle features in X-ray images, enabling more accurate and reliable classification.

While our current model exhibits exceptional results, there are inherent limitations. As with any machine learning model, it is crucial to recognize the boundaries of its capabilities. The accuracy achieved is not absolute, and there may be instances where misclassifications occur. Understanding these limitations is paramount for responsible deployment in medical contexts.

Looking forward, our focus revolves around continuous improvement. By incorporating a wider variety of X-ray images, we aim to ensure the model's adaptability to diverse patient demographics and anatomical variations, thereby fortifying its utility in clinical settings. The incorporation of interpretability tools such as Partition Explainer and SHAP values enhances the model's transparency, providing insights into decision-making processes.

In conclusion, our pursuit is anchored in advancing the classification of vertebrae X-ray images, contributing significantly to the medical field's diagnostic capabilities. As we navigate future developments, we remain dedicated to the responsible and progressive evolution of our model for the betterment of patient care and medical decision-making.

REFERENCES

[1] F. Zaina, R. Marchese, S. Donzelli, C. Cordani, C. Pulici, J. McAviney, and S. Negrini, "Current Knowledge on the Different Characteristics of Back Pain in Adults with and without Scoliosis: A Systematic Review," Journal of Clinical Medicine, vol. 12, no. 16, pp. 5182, 2023.

[2] Y. Sağlam, I. Bingöl, N. E. Yaşar, E. Dumlupınar, N. Ata, M. M. Ülgü, Ş. Birinci, G. Özdemir, O. Aslantürk, B. Görgün, et al., "The burden of scoliosis: A nationwide database study on demographics, incidence, and surgical rates," European Spine Journal, pp. 1-8, 2023.

[3] M. Karsy, A. K. Chan, P. V. Mummaneni, M. S. Virk, M. Bydon, S. D. Glassman, K. T. Foley, E. A. Potts, C. I. Shaffrey, M. E. Shaffrey, et al., "Outcomes and complications with age in spondylolisthesis: An evaluation of the elderly from the Quality Outcomes Database," Spine, vol. 45, no. 14, pp. 1000-1008, 2020.

[4]   Y. Ishimoto, C. Cooper, G. Ntani, H. Yamada, H. Hashizume, K. Nagata, S. Muraki, S. Tanaka, M. Yoshida, N. Yoshimura, et al., "Is radiographic lumbar spondylolisthesis associated with occupational exposures? Findings from a nested case control study within the Wakayama spine study," BMC Musculoskeletal Disorders, vol. 20, pp. 1-8, 2019.

[5]   Z. Jan, F. Ahamed, W. Mayer, N. Patel, G. Grossmann, M. Stumptner, and A. Kuusk, "Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities," Expert Systems with Applications, vol. 216, pp. 119456, 2023.

[6]   Y. A. Al-Naser, "The impact of artificial intelligence on radiography as a profession: A narrative review," Journal of Medical Imaging and Radiation Sciences, vol. 54, no. 1, pp. 162-166, 2023.

[7]   H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," BMC Medical Imaging, vol. 22, no. 1, pp. 69, 2022.

[8]   W. Wang, D. Liang, Q. Chen, Y. Iwamoto, X.-H. Han, Q. Zhang, H. Hu, L. Lin, and Y.-W. Chen, "Medical image classification using deep learning," in Deep Learning in Healthcare: Paradigms and Applications, pp. 33-51, Springer, 2020.

[9]   J. J. Thomas, A. A. Stans, T. A. Milbrandt, H. M. Kremers, W. J. Shaughnessy, and A. N. Larson, "Trends in incidence of adolescent idiopathic scoliosis: A modern US population-based study," Journal of Pediatric Orthopedics, vol. 41, no. 6, p. 327, 2021.

[10] P. Chen, Z. Zhou, H. Yu, K. Chen, and Y. Yang, "Computerized-assisted scoliosis diagnosis based on Faster R-CNN and ResNet for the classification of spine X-ray images," Computational and Mathematical Methods in Medicine, vol. 2022, 2022.

[11] J. Fatima, M. Mohsan, A. Jameel, M. U. Akram, and A. M. Syed, "Vertebrae localization and spine segmentation on radiographic images for feature-based curvature classification for scoliosis," Concurrency and Computation: Practice and Experience, vol. 34, no. 26, pp. e7300, 2022.

[12] S. Han, H. Zhao, Y. Zhang, C. Yang, X. Han, H. Wu, L. Cao, B. Yu, J.-X. Wen, T. Wu, et al., "Application of machine learning standardized integral area algorithm in measuring the scoliosis," Scientific Reports, vol. 13, no. 1, pp. 19255, 2023.

[13] W. Caesarendra, W. Rahmaniar, J. Mathew, and A. Thien, "AutoSpine-Net: Spine detection using convolutional neural networks for Cobb angle classification in adolescent idiopathic scoliosis," in Proceedings of the 2nd International Conference on Electronics, Biomedical Engineering,

and Health Informatics: ICEBEHI 2021, 3--4 November, Surabaya, Indonesia, pp. 547-556, Springer, 2022.

[14] F. Varçın, H. Erbay, E. Çetin, İ. Çetin, and T. Kültür, "End-to-end computerized diagnosis of spondylolisthesis using only lumbar X-rays," Journal of Digital Imaging, vol. 34, pp. 85-95, 2021.

[15] D. Saravagi, S. Agrawal, M. Saravagi, J. M. Chatterjee, M. Agarwal, et al., "Diagnosis of Lumbar Spondylolisthesis Using Optimized Pretrained CNN Models," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[16] F. Varçın, H. Erbay, E. Çetin, İ. Çetin, and T. Kültür, "Diagnosis of lumbar spondylolisthesis via convolutional neural networks," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1-4, 2019.

[17] 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, pp. e0267851, 2022.

[18] D. Saravagi, S. Agrawal, M. Saravagi, S. K. Jain, B. Sharma, A. Mehbodniya, S. Chowdhury, and J. L. Webber, "Predicting Lumbar Spondylolisthesis: A Hybrid Deep Learning Approach." Intelligent Automation & Soft Computing, vol. 37, no.2, pp. 2133–2151, 2023.

[19] A. Amin, M. Abbas, and A. A. Salam, "Automatic Detection and Classification of Scoliosis from Spine X-rays Using Transfer Learning," 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), pp. 1-6, 2022.

[20] A. A. A. Interiano, M. A. M. Palma, and K. M. R. Leiva, "Prediction of Spinal Abnormalities in Neuroradiology Images Applying Deep Transfer Learning," 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), pp. 1-7, 2023.

[21] D. Singh, J. Singla, M. K. I. Rahmani, S. Ahmad, M. ur Rehman, S. Jha, D. Prashar, J. Nazeer, et al., "Lumbar Spine Disease Detection: Enhanced CNN Model With Improved Classification Accuracy," IEEE Access, 2023

[22] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Progress in Artificial Intelligence, vol. 9, no. 2, pp. 85-112, 2020.

[23] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and MobileNet model for welding defects," IEEE Access, vol. 8, pp. 119951-119960, 2020.