IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 April 2024, accepted 18 April 2024, date of publication 8 May 2024, date of current version 29 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3398061

RESEARCH ARTICLE

Cervical Spine Fracture Detection and Classification Using Two-Stage Deep Learning Methodology

MUHAMMAD YASEEN[1], **MAISAM ALI**[1], **SIKANDAR ALI**[1], **ALI HUSSAIN**[1], **MOON-IL JOO**[1], **AND HEE-CHEOL KIM**[2]

[1]Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, South Korea
[2]College of AI Convergence, Institute of Digital Anti-Aging Healthcare, u-AHRC, Inje University, Gimhae 50834, South Korea

Corresponding author: Hee-Cheol Kim (heeki@inje.ac.kr)

This research was supported by the MSIT (Ministry of Science Information and Communication Technology), Korea, under the National Program for Excellence in Software (SW), supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863).

ABSTRACT Cervical spine fractures are a medical emergency that can cause permanent paralysis and even death. Traditional fracture detection techniques, such as manual radiography image interpretation, are time-consuming and prone to human error. Deep learning algorithms have shown promising results in various medical imaging applications i.e., disease diagnosis, including fracture detection of bones. In this study, we propose a two-stage approach for detecting cervical spine fractures. The first stage employs a convolutional neural network (CNN) model to determine the presence or absence of a fracture in the cervical spine, using a dataset of cervical spine Computed Tomography (CT) scan images as well as Grad-CAM for enhanced visualization and interpretation. In the second stage, our focus shifts to specific vertebrae within the cervical spine. To accomplish this task, we trained and evaluated the performance of the YOLOv5 and YOLOv8 models with 9170 images consisting of seven vertebrae. The detection results of both YOLO versions are compared and evaluated. The precision, recall, mAP50, and mAP50-90 were 0.900, 0.890, 0.935, 0.872, respectively. The results of this research demonstrate the potential of deep learning-based approaches for cervical spine fracture detection. By automating the detection process, these algorithms can assist radiologists and healthcare professionals in making accurate and timely diagnoses, leading to improved patient outcomes.

INDEX TERMS Cervical spine, computed tomography (CT), computer vision, deep learning, medical image, YOLO.

I. INTRODUCTION

Cervical spine, a crucial part of the human skeletal system, comprises of seven vertebrae stacked on top of each other, categorized as C1, C2, C3, C4, C5, C6, and C7 as shown in Figure 1. These vertebrae are separated by intervertebral discs and connected by ligaments. At its upper end, the cervical spine interfaces with the skull, while at the lower end, it links up with the thoracic spine. Traumatic incidents affecting the cervical spine led to a significant number of emergency department visits, with more than one million cases recorded annually in North America [1]. Cervical spine fractures and dislocations can result from severe trauma caused by events including falls, motor vehicle accidents, and sports injuries [2]. Common symptoms of cervical spine injury often include neck pain, which can move to other areas, such as the head, jaw, and shoulders. Additionally, individuals with such injuries may experience sensations of numbness and weakness in various parts of the body [3]. People commonly overlook neck pain as a common ailment caused by stress and work pressure. However, ignoring neck pain can have catastrophic effects on the cervical spine if not

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

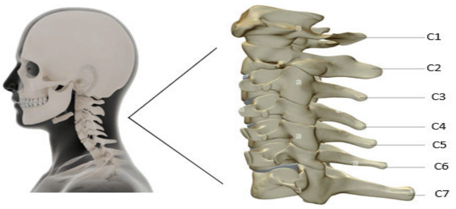

FIGURE 1. Anatomy of cervical spine.

treated promptly [4]. Patients engaged in accidents may be in an unstable condition in some situations, demanding rapid medical attention and vigilant monitoring to avoid delays in critical procedures. Rapid and simple diagnosis of fractures or injuries in the cervical spine can greatly speed up therapy for these unstable patients [5]. The prompt diagnosis of such injuries might be important in guaranteeing their well-being and avoiding additional difficulties. This study aims to develop deep learning models capable of diagnosing and detecting cervical spine fractures more accurately and quickly than traditional radiological methods.

A. CONVOLUTIONAL NEURAL NETWORK

The Convolutional Neural Network (CNN) is one of the most prominent and widely used neural network architectures in the realm of deep learning [6]. CNNs are feedforward neural networks that can extract features from data using convolution architecture. In contrast to traditional feature extraction methods [7], CNNs do not require manual feature extraction. CNNs possess the inherent ability to automatically learn and extract relevant features from raw data, eliminating the need for human intervention in the feature engineering process. In this model, we leverage deep convolutional neural network layers to extract meaningful features from CT scans of the cervical spine. The objective is to identify the presence of any fractures within the cervical spine based on these extracted features. The model utilizes a CNN architecture, featuring Conv2d layers for feature extraction, MaxPool2d layers for dimension reduction, ReLu layers for non-linearity, a Flatten layer to prepare data, and Linear layers for making predictions. This CNN structure efficiently analyzes CT scans of the cervical spine, aiding in fracture detection.

B. GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING (Grad-CAM)

Deep learning has demonstrated impressive achievements in the realm of medical image analysis. However, a significant challenge associated with deep learning models is their lack of interpretability, a critical concern within the medical field [8]. Deep learning models are constructed from neural networks, consisting of numerous layers of interconnected artificial neurons [9]. These models are trained on extensive datasets and have the capacity to discern intricate patterns and relationships. Nonetheless, they are often regarded as ''black boxes'' due to their lack of transparency in decision-making. Understanding the rationale behind their predictions or decisions can be challenging. Grad-CAM is a technique used to visually highlight the specific areas within an image that hold significance for a particular classification task. It's designed to provide insight into how a neural network, particularly a Convolutional Neural Network (CNN), arrives at its decisions when analyzing images.

C. YOLOv8 NETWORK

The YOLOv8 [10] object detection model is the most recent version. In January 2023, YOLOv8 was introduced by Ultralytics, the company that developed YOLOv5 [11]. This new iteration of the YOLO (You Only Look Once) series comes in five different scaled versions: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra-large). What makes YOLOv8 noteworthy is its versatility in supporting various computer vision tasks. These tasks encompass object detection, image segmentation, pose estimation, object tracking, and image classification. It has significant enhancements over previous versions of YOLO, such as a new neural network architecture that employs both the Feature Pyramid Network (FPN) and the Path Aggregation Network (PAN). The FPN reduces the spatial resolution of the input image gradually while increasing the number of feature channels. As a result, feature maps capable of detecting objects at various scales and resolutions are created. The PAN design, on the other hand, combines features from multiple network levels via skip connections. YOLOv8 adopts an anchor-free model strategy, featuring a decoupled head architecture. This design allows the model to handle objectness, classification, and regression tasks independently. The sigmoid function was utilized as the activation function for the objectness score in the YOLOv8 output layer, expressing the chance that the bounding box includes an object [12]. It uses the softmax function to express the probability of items belonging to each possible class. YOLOv8 may either be installed as a PIP package or executed through the command line interface (CLI). YOLOv8 represents a significant step forward in the field of computer vision, offering a range of models optimized for different applications and scales. There are two main objectives of this research study. First, the classification between the fractured and non-fractured spinal bones by leveraging the deep learning model. Second, the detection of fractured bones from seven different bone types using the state-of-the-art YOLO based object detection model. This proposed study performs both classification and detection. The detection of specific fractured bone types helps physicians to provide customized and better treatment and medication according to the nature of bone which also results in quick recovery from fracture.

The followings are our key contributions to this research study:

- Creation of bounding boxes from segmentation information for identification of cervical spine vertebrae.
- Introduction of a two-stage approach for detecting cervical spine fractures using deep learning algorithms.
- Proposed a deep learning model for determining the presence or absence of a fracture in the cervical spine

based on CT scan images, along with Grad-CAM for enhanced visualization and interpretation.

- Training and evaluation of YOLOv5 and YOLOv8 models for detecting specific vertebrae within the cervical spine, with comparison and evaluation of their detection results based on precision, recall, mAP50, and mAP50-90.

II. LITERATURE REVIEW

Deep learning is one of the fastest-growing areas of medical image analysis, and it has had a substantial impact on a variety of clinical and scientific applications. New breakthroughs are paving the way for a bright future for artificial intelligence, paving the way for more accurate segmentation, classifications, detections, and forecasts even at the level of professional radiologists. Deep learning approaches outperform traditional methods in medical image analysis. But there is still a long way to go.

Salehinejad et al. [13] developed a deep sequential learning approach for identifying injuries and fractures in the cervical region of the spine using computed tomography. In this research, the authors designed a deep convolutional neural network (CNN) that incorporates recurrent layers with bidirectional capabilities. This architecture was employed for the automatic identification and detection of fractures in axial CT images of the spine. The validation of the data resulted in classification accuracies of 70.92% and 79.18%. The primary limitation of this approach is its tendency to produce a high number of false positives, which can result in inaccurate predictions across various cases. When an image within a single case is classified as a false positive, it can influence the entire case, causing it to be classified as a false positive. These results highlight the importance of including common features in the training process. The high performance on imbalanced datasets is primarily attributed to the dataset's bias towards negative cases and images.

Erickson et al. [14] mentioned that deep learning is a popular approach for performing a range of crucial tasks in radiology and medical imaging. Some types of deep learning can precisely segment organs (trace the borders, allowing volume measurements or other attributes to be calculated). Other deep learning networks can predict significant qualities from image areas, such as whether something is cancerous, molecular markers for tissue in a region, and even prognostic indicators. Deep learning is simpler to train than classic machine learning approaches, but it requires more data and greater caution when interpreting results. It will automatically identify the key elements, but it might be a challenge to identify what those features are. The preceding study reveals in detail how precise, popular, and significant deep learning models are. Additionally, this article discusses the fundamental ideas of deep learning systems as well as some of the pitfalls that can occur while developing deep learning systems and how to avoid them.

Chłąd and Ogiela [15] explore the utilization of deep learning and cloud-based computing in the development of a system for the detection of cervical spine fractures. The study evaluates the effectiveness of vision transformers (ViT) in detecting fractures and proposes a cloud-based system for automatic inference and training of such models. The results show that ViT models achieve high accuracy for the classification of fractures. Highlights the importance of augmentation procedures in improving the performance of vision transformers. Cloud-based systems enable the implementation of large-scale detection systems, improving the efficiency of medical personnel. Overall, vision transformers offer a promising approach to enhance fracture detection in the cervical spine. However, there is still a challenge in accurately detecting fractures in specific vertebrae.

Krawczyk and Starzynski [16] The article primarily focuses on utilizing the You Only Look Once (YOLO) neural network to identify individual bones within a series of CT image slices depicting the human pelvic area. To achieve this, the YOLO network underwent training using custom data, enabling it to recognize and locate various bone structures in CT images. Subsequently, the network was applied to a different set of CT data to assess its accuracy in detecting bone structures. In the final step, bounding boxes generated by the YOLO algorithm for the detected regions were used to position idealized bone models within the CT data, facilitating a precise alignment of these models with the actual bone structures observed in the CT scans.

III. MATERIALS AND METHODS

This section will provide a detailed explanation of the materials used and the methodology employed for the development of a deep learning-based system for cervical spine fracture detection.

A. DATASET DESCRIPTION

The RSNA 2022 Cervical Spine Fracture Detection Challenge dataset [17] served as the basis of this study. This comprehensive dataset comprises a collection of 2019 CT scans specifically focused on the cervical spine area. Each individual scan encompasses a range of 100 to 800 slices, each varying in thickness.

One of the notable features of this dataset is the inclusion of bounding boxes within a subset of the training set. These bounding boxes serve as indicators of areas of interest, particularly regions that might have sustained damage. It's worth noting, however, that most slices in the dataset lack such annotations. Consequently, this absence of annotations poses a challenge in determining whether a particular slice showcases any signs of damaged vertebrae.

The availability of segmentations for portions of the scans is an intriguing feature of the dataset. These segmentations have been automatically generated using a 3D U-Net model. Subsequently, experienced radiologists meticulously reviewed and modified these segmentations to ensure accuracy [17]. The segmentation labels provided encompass a range of values from 1 to 7, representing the seven cervical

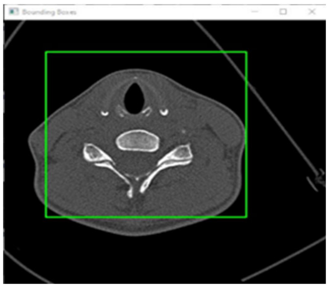

FIGURE 2. Sample image with bounding box.

vertebrae from C1 to C7. Notably, a label of 0 designates the background region.

B. DATA PREPARATION

We initiated data preparation with the conversion of DICOM images into the PNG format, a critical step in our workflow. To optimize the visual representation of these images, we employed windowing techniques, specifically setting the window center to 500 and the window width to 2000. This strategic choice resulted in a transformation of the image's appearance, highlighting specific anatomical structures of interest with enhanced clarity [18], [19], [20]. These adjustments contributed to an overall improvement in the visual quality of the images, a crucial aspect of our data preparation process.

Following the data preparation phase, the next step in our workflow involved preparing the data for use in our classification models. These models rely on the availability of labeled images, where each image is classified as either displaying a positive or negative fracture in the cervical spine area. To establish these labels, we leveraged slices containing bounding box annotations, designating them as positive cases. Conversely, slices lacking bounding box annotations were considered negative instances [15].

This approach yielded a dataset of approximately 7217 images classified as positive cases. However, a notable challenge emerged due to the inherent class imbalance within the dataset. To rectify this imbalance and create a more equitable distribution, we employed an under-sampling technique. This technique involves the deliberate reduction of instances in the overrepresented class, in this case, the negative class. By implementing under-sampling, we aimed to mitigate the potential bias that could arise from an imbalanced dataset.

The process of under-sampling involved the random selection of negative slices from the pool of available negative instances. This strategic sampling helped balance the number of positive and negative instances in the dataset, ultimately enhancing the training process of our classification models [21 22]. The incorporation of under-sampling contributed to a more reliable and representative training set, enabling our models to learn and generalize effectively across both positive and negative classes.

In pursuit of detecting cervical spine bone, we leveraged the segmentation information available in the dataset.

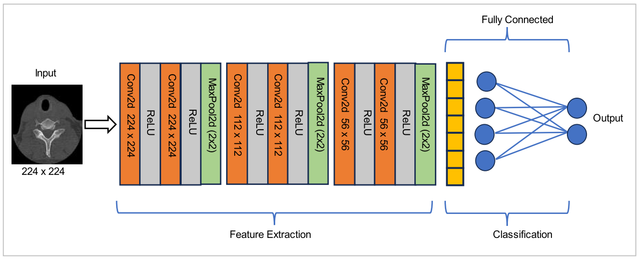

FIGURE 3. Architecture of proposed deep learning model.

This segmentation delineated the distinct cervical vertebrae, assigning each vertebra a unique numerical label, ranging from 1 to 7, along with a background label denoted as 0. The subsequent step involved the creation of bounding boxes. Utilizing the segmentation information, we accurately outlined bounding boxes around each cervical vertebra, shown in Figure 2. These bounding boxes effectively encapsulated the regions of interest within the images.

Crucially, we assigned labels to the slices in correspondence with the cervical spine number. This means that each image was associated with the numerical label of the cervical vertebra based on the segmentation information. This process imbued each image with a clear and meaningful label, enabling the training of the YOLOv8 model for accurate detection of cervical spine bone.

C. METHODOLOGY

This section outlines the implementation of deep learning models within the scope of this study. It discusses how various neural network architectures were utilized, their configurations, and the techniques applied to train and evaluate these models.

1) CLASSIFICATION

In the classification process, we initiated with a crucial step of resizing the images, ensuring uniform dimensions of 224 × 224 pixels. Subsequently, we divided the dataset into training and validation subsets, carefully allocating them in an 80:20 ratio to facilitate model training and evaluation. We began data augmentation with the core data structures in place, using parameters such as random horizontal flip, random rotation, Gaussian blur, and more. This augmentation not only enriched the dataset but also strengthened the potential accuracy of our models, enhancing their ability to make precise classifications.

To classify whether bones were fractured or not, we employed a variety of models, including DenseNet, ResNet, and a proposed deep learning model. These models were utilized to analyze and make predictions regarding the presence or absence of fractures in bone structures. The proposed deep learning model commences with Conv2d layers that perform essential convolution operations with varying numbers of input and output channels, while a Rectified Linear Unit (ReLU) activation function follows each convolution for introducing non-linearity. MaxPool2d layers are

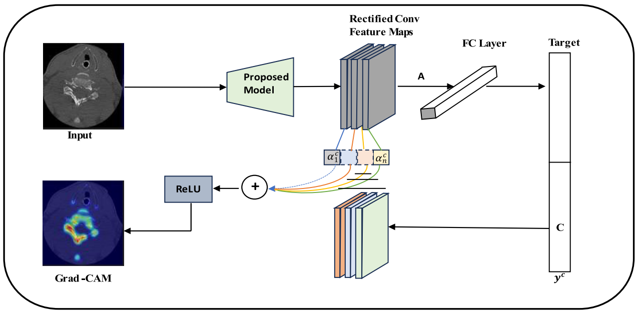

FIGURE 4. Grad-CAM's overview.

introduced to carry out max-pooling, which serves to reduce the spatial dimensions of the feature maps, retaining critical information. After the convolutional operations, a Flatten layer reconfigures the 2D feature maps into a 1D vector, preparing the data for fully connected layers. Subsequently, a series of Linear layers, equipped with ReLU activations, engage in the final classification task. The first Linear layer produces 512 output features, followed by another with 128 output features, and ultimately, a Linear layer with 2 output features. These two outputs likely correspond to the two classes that the model aims to distinguish, such as "fracture" and "non-fracture." In essence, this model takes input images, extracts salient features through convolutions, and leverages fully connected layers for the ultimate classification into one of the two classes. Figure 3 shows the architecture of our proposed model.

The proposed model has several layers, including convolutional, max pooling, and fully connected layers. The model uses input image tensors X to generate output predictions. The mathematical form of the model can be described as follows:

a: CONVOLUTIONAL LAYER

The output of each convolutional layer l is obtained by applying a convolution operation followed by a Rectified Linear Unit (ReLU) activation function:

$$X^{(l+1)} = ReLU(Conv2d\left(X^{(l)}, W^{(l)}, b^{(l)},\right. \left. kernel_{size} = K^{(l)}, padding = p^{(l)}\right)) \quad (1)$$

b: MAX POOLING LAYER

$$X^{(l+1)} = MaxPool2d(X^{(l)}, kernel_{size} = 2, stride = 2) \quad (2)$$

c: FLATTEN LAYER

$$X^{(l+1)} = Flatten(X^{(l)}) \quad (3)$$

Here, $X^{(l)}$ denotes the input tensor of layer l while $X^{(l+1)}$ represents the output tensor of the same layer. $W^{(l)}$ and $b^{(l)}$ refer to the weights and biases of layer l, respectively. The kernel size of layer l is denoted by $K^{(l)}$ and the padding size by $p^{(l)}$. Operations such as convolution (*Conv2d*), max pooling (*MaxPool2d*), and flattening (*Flatten*) are employed to process the input tensor.

However, in recognition of the critical necessity of interpretability alongside accuracy in medical decision-making, we went beyond quantitative measurements. Incorporating Grad-CAM visualization to our developed model was instrumental in shedding light on the inner workings of the models' decision processes [23]. By strategically applying Grad-CAM to the final convolutional layers of each model, we could discern the specific regions within the images that significantly influenced their predictions [24]. Implementation of Grad-CAM begins by defining the target for Grad-CAM and specifying the last convolutional layer of the proposed deep learning model. Grad-CAM is then applied to generate grayscale Class Activation Maps (CAMs) based on an input image and target predictions. These grayscale CAMs are transformed into colorized CAM images by overlaying them onto the original image. The final result is a side-by-side comparison of the original image and colorized CAM, providing insights into which areas of the image influence the model's classification decision. This technique aids in understanding the neural network's decision-making process, especially in complex deep-learning models.

Figure 4 shows an overview of Grad-CAM. It does this by examining the gradients of the model's prediction with respect to the image's pixel values. Essentially, it highlights which areas of the input image had the strongest influence on the model's final decision, helping to make the neural network's decision-making process more interpretable [25], [26]. Grad-CAM computes the gradient of y^c (score for class c) with respect to feature mappings A of a convolutional layer to get the class-discriminative localization map. Equation (4) is used to compute the importance or relevance of each spatial location in an input image for a specific class prediction.

$$\alpha_k^c = \frac{1}{Z}\sum\nolimits_i \sum\nolimits_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (4)$$

where:

- α_k^c represents the importance or contribution of the spatial location k in the feature map A to the score of class c.
- Z is a normalization factor.
- The double summation $\Sigma_i \Sigma_j$ implies summing over the spatial dimensions of the feature map.
- $\frac{\partial y^c}{\partial A_{ij}^k}$ is the partial derivative of the score for class c with respect to the activation at spatial location (i, j) in the feature map A^k.

The Grad-CAM heat-map is a weighted combination of feature maps, however it is preceded by a ReLU. The ReLU activation is applied to focus on regions that positively contribute to the prediction of class c.

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (5)$$

where, $L_{Grad-CAM}^c$ represents the Grad-CAM localization map for class c.

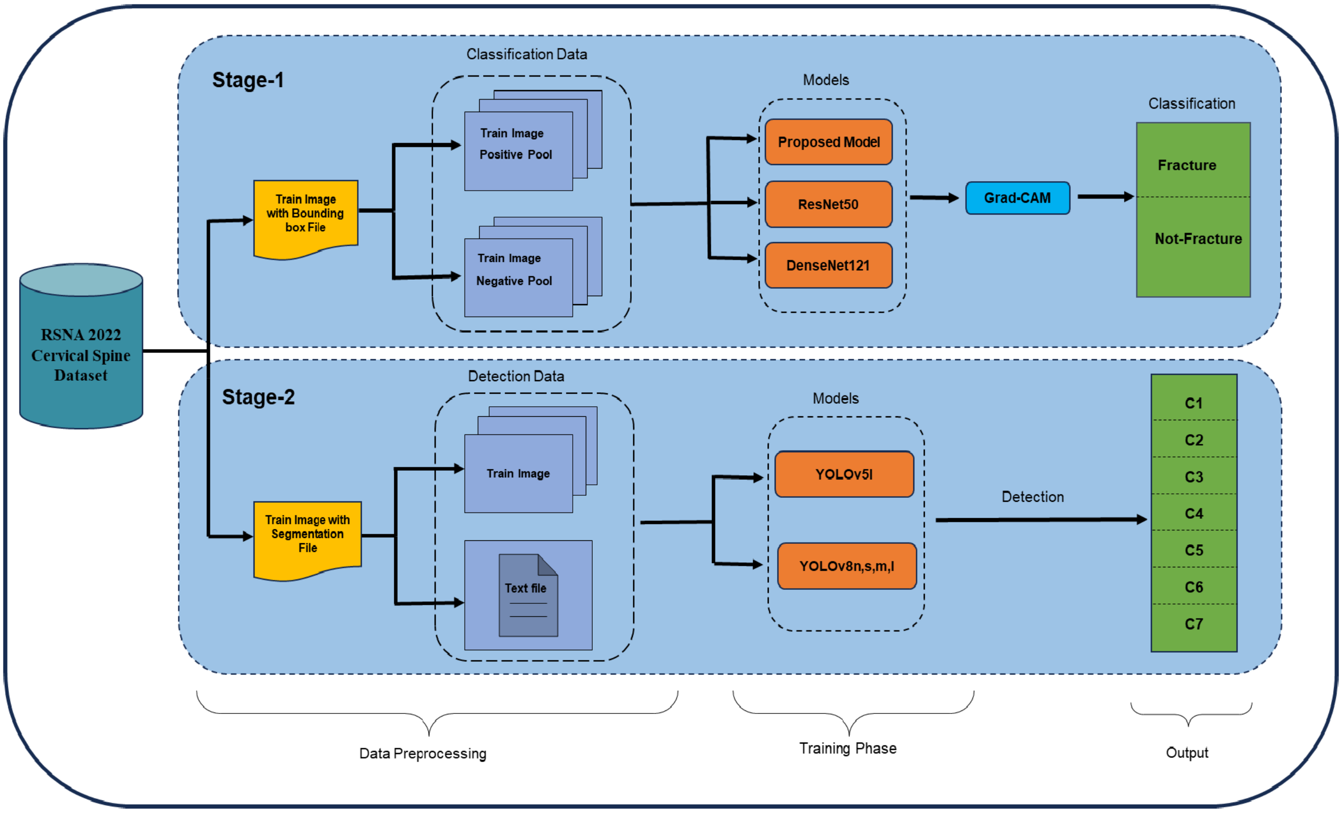

FIGURE 5. The overall workflow of our proposed study.

TABLE 1. Training parameters of YOLOv8.

Parameters	Values
Batch size	16
Image size	512
Learning rate	0.01
Epochs	200
Optimizer	SGD
Momentum	0.937
Weight decay	0.0005

2) DETECTION

The cervical spine consists of seven vertebrae labelled as C1 to C7. Identifying the exact vertebrae being scanned is crucial to pinpointing the specific area of the spine and classifying any fractures presence. To accomplish this, we employed a cutting-edge model called YOLOv8. It is critical to select the best training parameters for each version of YOLO in order to acquire the best object detection results. Among the most crucial training parameters are learning rate, the number of training iterations, and batch size. The learning rate influences how rapidly the network learns from new information from training data; a greater learning rate may lead to instability or overshooting the optimal outcome. The number of training iterations establishes how often the network encounters the training data throughout the training process. Increasing the number of iterations usually leads to enhanced performance but may also raise the risk of overfitting or extending the training duration. The batch size specifies the number of images processed concurrently in each training step. Larger batch sizes typically lead to quicker training but could potentially lead to memory limitations or slower convergence. Apart from these parameters, every YOLO version might possess its unique set of optimal training parameters. We trained five YOLO versions and compared them. The dataset includes 9170 images, and 80% of the data was used to train the model. The image size was 512 × 512, and we selected a batch size of 16. The network was optimized using the SGD optimizer with a learning rate of 0.01. The hyperparameters were selected through a random search approach, where various combinations of hyperparameters were sampled randomly. These combinations were then evaluated based on the model's accuracy and performance metrics. The set of hyperparameters that yielded the highest accuracy during this evaluation process was chosen for further experimentation and analysis. This method allowed for an empirical determination of the optimal hyperparameter values based on their impact on the model's predictive performance. The model was trained using the train.py script, which was executed while specifying the batch size, image size, weights of the specific model, epochs, etc. Some folders must be maintained in order to train the model. Divide the dataset folder into two folders: one for the labels, and one for the images. All the images are in the image folder, while the labels are in the label folder. The labels are text files that provide information on the object classes, bounding boxes, coordinates, height, and width of the bounding boxes, as well as the labels for each image. In this research, we chose the training parameters outlined in Table 1 to ensure optimal performance. By using YOLOv8, we aimed to achieve accurate and efficient identification of the particular vertebrae under examination, contributing to improved diagnostic capabilities in cervical spine fracture detection.

Overall, we proposed an approach for the detection of cervical spine fracture, consisting of two main stages. In the first stage, we employ a proposed deep learning model to determine whether a fracture is present or not. This classification helps us decide if there's a fracture in the cervical spine. We also utilize a visualization technique called Grad-CAM to understand better how the model makes its decisions. Moving to the second stage, we focus on detecting the specific vertebrae within the cervical spine. We use the YOLOv8 model for this task, which is highly proficient in recognizing and localizing objects. By implementing YOLOv8, we aim to identify the individual vertebrae precisely, contributing to a comprehensive approach in our study of cervical spine fracture detection. Figure 5 provides a comprehensive visual representation of the entire methodology employed in this study, offering an overview of the various steps, techniques, and processes used in the research.

D. PERFORMANCE EVALUATION

The model's evaluation involved the use of various crucial performance metrics, including precision (P), recall (R), F1-score (F1), Cohen's Kappa (K), Specificity (Sp) and mean average precision (mAP), as these metrics play a pivotal role in assessing the effectiveness of the object detector. The specific equations for these performance metrics are presented below.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{8}$$

1) COHEN'S KAPPA

Cohen's Kappa, also commonly referred to as "kappa," is a statistic in machine learning that assesses the degree of agreement between two raters (or classifiers) when classifying items into groups that are mutually exclusive. It is particularly useful when dealing with classification tasks. The value of Kappa ranges from −1 to 1, where 1 indicates perfect agreement, 0 indicates agreement due to chance, and −1 indicates complete disagreement.

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{9}$$

where:

- p_0 is the observed proportion of agreement between raters.
- p_e is the expected proportion of agreement between raters if the agreement occurred by chance.

2) SPECIFICITY

Specificity is a statistical measure used to evaluate the performance of a binary classification model, particularly in medical diagnosis or screening tests. It measures the ability of the model to correctly identify true negatives, or the proportion of actual negative cases that are correctly classified as negative by the model.

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

where:

- TN represents True Negatives (the number of negative instances correctly classified as negative).
- FP represents False Positives (the number of negative instances incorrectly classified as positive).

$$mAP = \frac{1}{N} \sum_{N}^{i=1} AP_i \tag{11}$$

where TP, FP and FN are the numbers of true positives, false positives and false negatives respectively, represent the number of instances detected by model. The mean average precision (mAP) is calculated by taking the average precision of all classes and the overall intersection over union (IoU) threshold. Precision measures how accurate the model's prediction is, whereas recall measures how well the model recognizes all positives. The mean average precision (mAP) is a well-known statistic for evaluating the performance of object detectors. It computes the average precision value for recall values ranging from 0 to 1.

E. EXPERIMENTAL ENVIRONMENT

All experiments, data pre-processing, and analysis were executed on a 64-bit Windows operating system equipped with a 12th Gen Intel(R) Core (TM) i7-12700 processor operating at 2.10 GHz, and 64 GB of installed RAM. The experiments and model training utilized cuda version 11.8 with an RTX 4060 GPU. The software stack involved Pytorch, OpenCV, Python (Version 3.9), and Scikit-learn libraries, collectively supporting the research and experimentation process.

IV. EXPERIMENTAL RESULTS

This section provides an overview of the experimental results pertaining to both the classifier and detection components of our study.

A. CLASSIFICATION RESULTS

Models were trained with training data and evaluated with validation data. Table 2 presents a comprehensive overview of the performance of various deep learning models. To assess

TABLE 2. Comparative analysis of classifiers.

Classifier	P %	R %	F1 %	Acc %	K %	Sp %
DenseNet121	96.4	91.7	94.0	94.2	92.6	96.5
ResNet50	96.8	92.8	94.7	94.9	93.8	97.0
Proposed Model	**98.1**	**97.5**	**97.8**	**97.8**	**96.6**	**99.0**

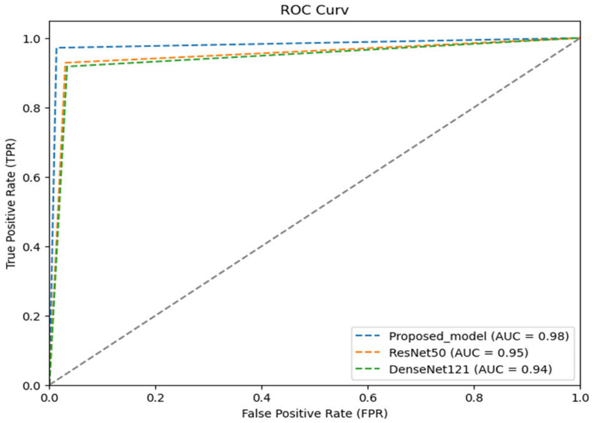

FIGURE 6. ROC curve of algorithms.

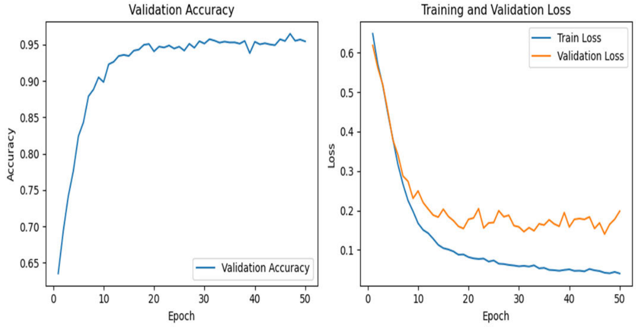

FIGURE 7. Accuracy and loss graph of proposed model.

the output results of these models, we employed evaluation techniques, including precision, recall, F1-score, and accuracy. Notably, the accuracy scores for ResNet50, DenseNet121, and proposed deep learning model are 94.2%, 94.9%, and 97.8%, respectively.

The Receiver Operating Characteristic (ROC) curve serves as graphical representations of model performance. Figure 6 illustrates the graphical representation of different models, offering comprehensive insights into their performance characteristics.

In Figure 7, the graph depicts the training and validation accuracy and loss throughout the training phase of the selected model. As the number of epochs increases, there is a noticeable trend of accuracy rising and loss diminishing, indicating the model's improvement over time.

Figure 8 displays the confusion matrix of the selected model. The analysis of the confusion matrix offers a comprehensive perspective on the model's performance, with a particular focus on the main diagonal values, which provide insights into the model's accuracy. This matrix is a valuable tool for understanding the model's ability to correctly classify and predict classes.

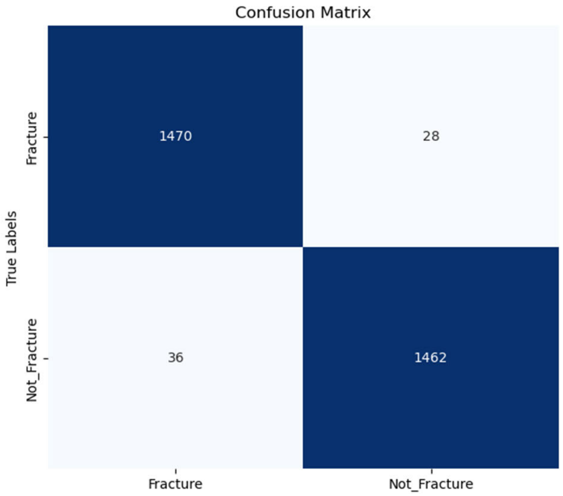

FIGURE 8. Confusion matrix of proposed model.

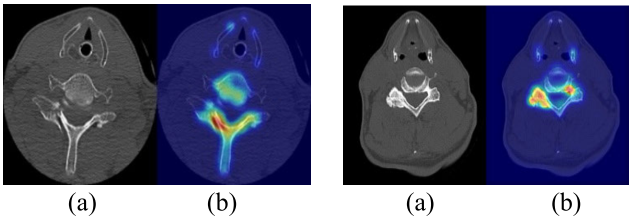

FIGURE 9. Output results of Grad-CAM. Figure (a) shows the input image while image (b) shows the Grad-CAM output for the two example images.

B. GRAD-CAM RESULTS

The proposed model's decisions are validated and visually represented through the utilization of Grad-CAM. Figure 9 provides a visual representation of the Grad-CAM process. In this context, (a) represents the input image and (b) illustrates the output image produced by Grad-CAM, which highlights the regions within the input image that were instrumental in the proposed model's decision. This visualization aids in understanding the key areas that influenced the model's classification, making it a valuable tool for enhancing the interpretability and transparency of the model's decisions in the context of medical image analysis.

C. DETECTION RESULTS

We applied YOLOv5l and all four versions of YOLOv8, namely YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l. The findings were compared, and the experimental results demonstrated that YOLOv8l had a greater recognition rate than the other models. Table 3 compares the different YOLO architectures, and the results are displayed by analyzing numerous metrics such as precision, recall, mAP50, and mAP50-95.

TABLE 3. Comparative analysis of different YOLO versions.

YOLO Versions	Precision %	Recall %	mAP50 %	mAP50-90 %
YOLOv5l	86.2	88.7	91.9	80.3
YOLOv8n	88.2	88.2	92.6	82.3
YOLOv8s	88.9	88.7	92.6	85.1
YOLOv8m	88.2	90.9	93.4	86.4
YOLOv8l	**90.0**	**89.0**	**93.5**	**87.2**

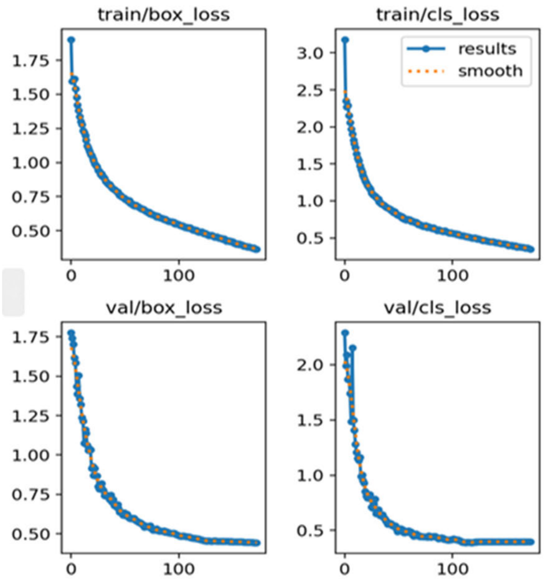

FIGURE 10. Shows the box loss and class loss for the training and validation, respectively.

In Figure 10, box loss and class loss are shown for the training and validation phase. Box loss describes the error in estimating the size and location of the bounding boxes around the bones. By comparing the predicted bounding boxes and the actual bounding boxes for each bone, this loss is computed. Better precision in locating the bones is indicated by a lower box loss. However, class loss refers to the error in identifying the proper class or kind of bone for each bounding box. The anticipated class probabilities and the actual class labels are compared to determine the loss. Better accuracy in determining the proper type of bone is indicated by a smaller class loss. Figure 10 shows that both box loss and class loss decrease continuously as the epoch increases.

In Figure 11, we can observe the precision and recall values trends throughout the training process. Initially, the precision value started at a modest 0.263 but steadily increased over time, ultimately reaching an impressive level of 0.900. Similarly, the recall value began at 0.474 and showed a rapid ascent in the graph, achieving a noteworthy value of 0.800. It's noteworthy that during the initial training epochs, there was some fluctuation in the recall graph, indicating some variability in the model's ability to identify true positive cases. However, as the training progressed, the recall graph stabilized, and the model consistently achieved an impressive recall value of 0.890. Furthermore, the metrics mAP50 and mAP50-95, which are essential measures of model performance, were calculated at 0.935 and 0.872, respectively. These metrics underscore the model's exceptional performance, particularly in terms of accurately detecting and localizing bones within the dataset.

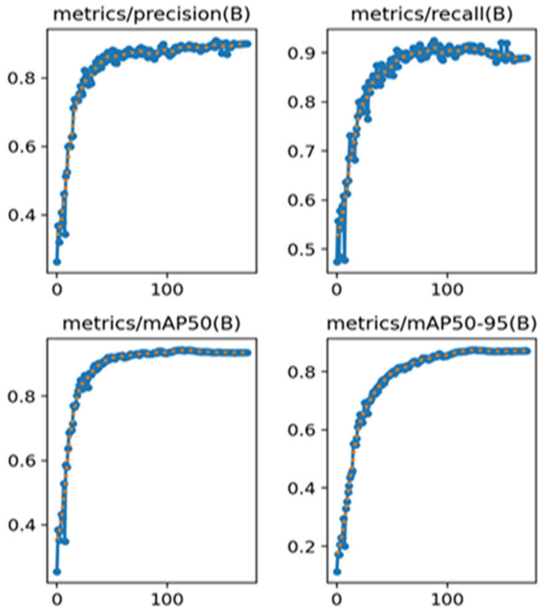

FIGURE 11. Shows precision, recall, mAP50, and mAP50-95 of YOLOv8l.

In Figure 12. confusion matrix shows the number of true positives (correctly identified vertebrae), true negatives (correctly identified wrong-vertebrae), false positives (incorrectly identified vertebrae), and false negatives (incorrectly identified wrong-vertebrae) for each vertebra in the context of detecting cervical spine vertebrae. The accuracy score for each vertebra represents the percentage of correct identifications out of all vertebral identifications. For example, an accuracy score of 91 for C1 suggests that the model successfully identified 91% of the C1 vertebrae in the dataset.

Figure 13 provides compelling evidence of the YOLOv8l model's exceptional performance in detecting cervical spine bones with remarkable confidence and precision. The results prominently showcase the model's ability to achieve confidence scores of over 90% in most instances. These outcomes not only underscore the model's outstanding performance but also demonstrate its remarkable accuracy in detecting objects within the images. The model consistently generates precise bounding boxes and confidence scores for each prediction, affirming its reliability and efficacy in bones detection tasks.

D. ABLATION STUDIES

This section presents the results of ablation studies conducted on our proposed model for classification and detection tasks. The objective of these experiments was to investigate the impact of different model components on performance

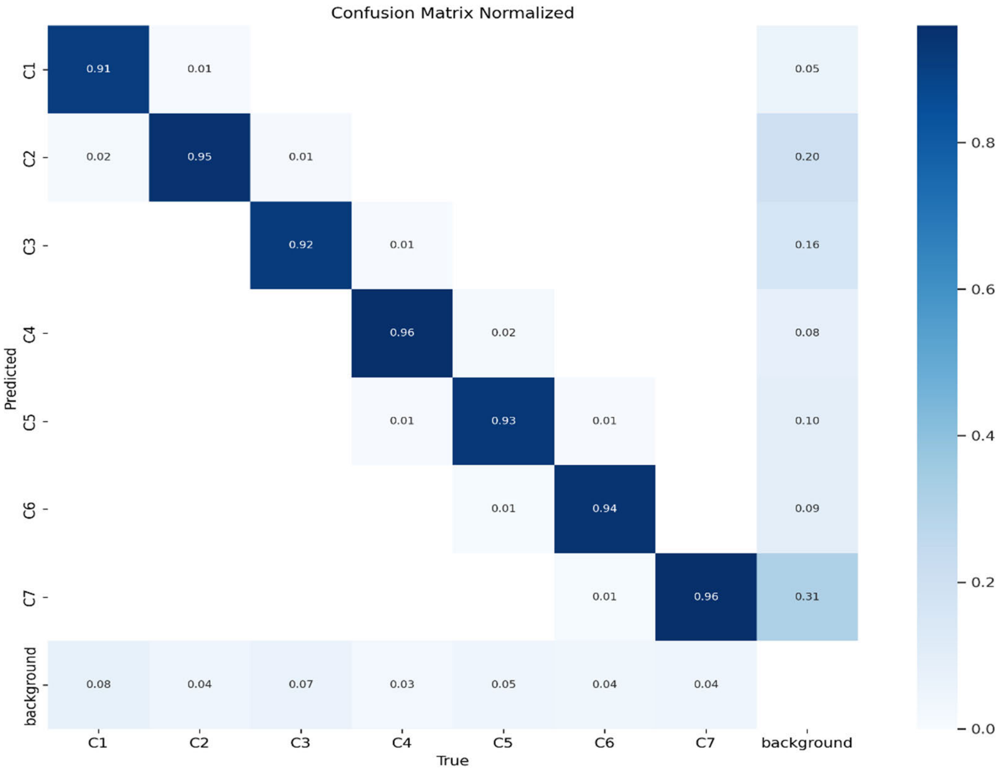

FIGURE 12. Confusion matrix of YOLOv8l.

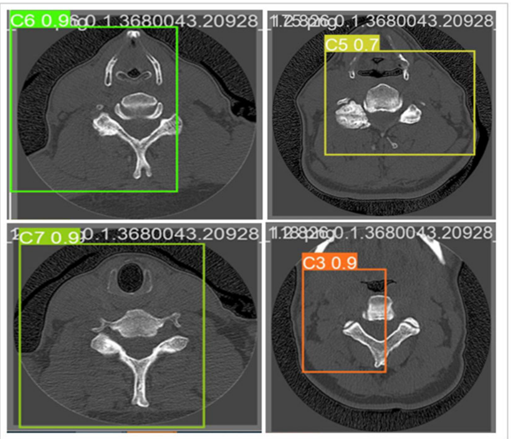

FIGURE 13. Detection of cervical spine bones by the YOLOv8 model.

metrics, specifically focusing on the effectiveness of fully connected layers for classification and the complexity of the YOLO model for object detection.

For the classification task, we conducted ablation studies to assess the contribution of fully connected layers within our proposed classification model. The experiments involved systematically varying the number of fully connected layers and evaluating their effects on classification accuracy. Specifically, we experimented with models containing 1 layer, 2 layers, and 3 layers of fully connected components. The results of these experiments, as summarized in Table 4, indicate that fully connected layers play a significant role in enhancing the effectiveness of our classification model. Interestingly, we observed that increasing the number of fully connected layers led to improvements in classification accuracy. Particularly, the model with three fully connected layers demonstrated the highest performance, suggesting that deeper connections within the network contribute to better classification outcomes.

TABLE 4. Results of ablation study on fully connected layers.

Fully Connected Layers	P %	R %	F1 %	Acc %
1 Layer	96.4	87.1	91.5	92.2
2 Layers	98.2	91.1	94.5	94.6
3 Layers	**98.1**	**97.5**	**97.8**	**97.8**

For detection focused on assessing the impact of model complexity. Through these experiments, we aimed to understand how variations in model complexity affect detection accuracy. Our findings reveal that as the complexity of the YOLO model increases, there is a corresponding improvement in detection accuracy as shown in Table 3. This underscores the importance of considering model complexity as a critical factor in designing effective object detection systems.

V. DISCUSSION

This study began with the collection and preprocessing of CT scan images, ensuring that they met the specific requirements of classification and detection algorithms. The use of CT scans is critical since it is a popular and reliable method for identifying cervical spine fractures [27], [28]. The preprocessing step is crucial to optimize the input data for deep learning models by normalizing pixel values, resizing images, and ensuring data consistency. This preparation was essential in achieving accurate and efficient results.

For fracture classification, we employed three distinct deep learning models: DenseNet121, ResNet50, and a custom sequential CNN. Each of these models exhibited impressive accuracy rates of 94.2%, 94.9%, and 97.8%, respectively. These high accuracy rates reflect the effectiveness of deep learning models in distinguishing between normal and fracture cases, which is crucial in medical diagnostics. The custom sequential CNN, in particular, achieved outstanding results, underlining the significance of specialized model architectures in enhancing classification accuracy.

To gain insights into the decision-making processes of our classification models and ensure transparency and interpretability, we applied Grad-CAM [25], [29], [30]. The Grad-CAM technique allowed us to visualize the regions within the CT scan images that the models focused on when making their fracture classification decisions. This visualization not only aids in understanding the model's reasoning but also holds promise for educational and clinical use, enabling healthcare professionals to verify and validate the model's diagnoses.

While the classification algorithms excelled in accurately categorizing fractures, a vital challenge emerged in the subsequent step - localization. The need to precisely determine the specific location of cervical spine fractures, particularly the vertebrae (C1, C2, . . . C7), was a complex task. This necessitated the integration of YOLOv8, a state-of-the-art object detection framework, to precisely identify and locate the specific parts of the cervical spine. The results of YOLOv8-based bone detection are impressive, with precision, recall, mAP50, and mAP50-90 scores of 90.0%, 89.0%, 93.5%, and 87.2%, respectively. These findings demonstrate the robustness and efficacy of YOLOv8 in accurately localizing cervical spine components, thus facilitating targeted and personalized medical interventions.

The implications of our findings are substantial, as they demonstrate the potential of deep learning models in cervical spine fracture diagnosis. By achieving high classification accuracy and effective fracture localization, our research holds the promise of improving the diagnostic process for healthcare professionals. The visual interpretability afforded by Grad-CAM enhances the trustworthiness of the models' decisions. Moreover, the high precision and recall rates in fracture detection highlight the potential for enhancing patient care, particularly in emergency situations.

VI. CONCLUSION

In conclusion, this research showcases the successful application of deep learning models, image preprocessing, visualization, and detection techniques in the context of cervical spine fracture diagnosis. The achieved accuracy rates, coupled with the ability to localize fractures, are instrumental in enhancing patient care, reducing complications, and expediting medical interventions. The findings of this study contribute to the growing body of knowledge in medical image analysis and have the potential to revolutionize cervical spine fracture diagnostics. However, the application of these techniques requires rigorous validation on a larger and more diverse data set, ensuring that the technology performs reliably under real-world conditions.

REFERENCES

[1] J. A. G. Crispo, L. J. W. Liu, V. K. Noonan, N. P. Thorogood, B. K. Kwon, M. F. Dvorak, D. Thibault, A. W. Willis, and J. J. Cragg, "Nationwide emergency department visits for pediatric traumatic spinal cord injury in the United States, 2016–2020," *Frontiers Neurol.*, vol. 14, pp. 2016–2020, Nov. 2023, doi: 10.3389/fneur.2023.1264589.

[2] C. Zanza, G. Tornatore, C. Naturale, Y. Longhitano, A. Saviano, A. Piccioni, A. Maiese, M. Ferrara, G. Volonnino, G. Bertozzi, R. Grassi, F. Donati, and M. A. A. Karaboue, "Cervical spine injury: Clinical and medico-legal overview," *La Radiologia Medica*, vol. 128, pp. 103–112, Jan. 2023, doi: 10.1007/s11547-022-01578-2.

[3] A. M. Torlincasi. (Aug. 22, 2022). *Cervical Injury*. StatPearls—NCBI Bookshelf. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK448146/

[4] A. A. Drosos, E. Pelechas, A. N. Georgiadis, and P. V. Voulgari, "A not-to-miss cause of severe cervical spine pain in a patient with rheumatoid arthritis: A case-based review," *Medit. J. Rheumatol.*, vol. 32, no. 3, p. 256, 2021, doi: 10.31138/mjr.32.3.256.

[5] M. B. S. Bhavya, M. V. Pujitha, and G. L. Supraja, "Cervical spine fracture detection using pytorch," in *Proc. IEEE 2nd Int. Conf. Mobile Netw. Wireless Commun. (ICMNWC)*, Dec. 2022, pp. 1–7, doi: 10.1109/ICMNWC56175.2022.10031629.

[6] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[7] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Jun. 2018, doi: 10.1007/s13244-018-0639-9.

[8] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is grad-CAM explainable in medical images?" 2023, *arXiv:2307.10506*.

[9] Y. Bengio and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://books.google.ie/books?id=Np9SDQAAQBAJ&printsec=frontcover&dq=Goodfellow,+I.,+Bengio,+Y.,+%26+Courville,+A.+(2016).+Deep+learning.+MIT+press.&hl=&cd=1&source=gbs_api

[10] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *YOLO By Ultralytics*. [Online]. Available: https://github.com/ultralytics/

[11] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 2023, *arXiv:2304.00501*.

[12] X. Zhai, Z. Huang, T. Li, H. Liu, and S. Wang, "YOLO-drone: An optimized YOLOv8 network for tiny UAV object detection," *Electronics*, vol. 12, no. 17, p. 3664, Aug. 2023, doi: 10.3390/electronics12173664.

[13] H. Salehinejad, E. Ho, H.-M. Lin, P. Crivellaro, O. Samorodova, M. T. Arciniegas, Z. Merali, S. Suthiphosuwan, A. Bharatha, K. Yeom, M. Mamdani, J. Wilson, and E. Colak, "Deep sequential learning for cervical spine fracture detection on computed tomography imaging," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1911–1914, doi: 10.1109/ISBI48211.2021.9434126.

[14] B. J. Erickson, P. Korfiatis, T. L. Kline, Z. Akkus, K. Philbrick, and A. D. Weston, "Deep learning in radiology: Does one size fit all?" *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 521–526, Mar. 2018, doi: 10.1016/j.jacr.2017.12.027.

[15] P. Chłąd and M. R. Ogiela, "Deep learning and cloud-based computation for cervical spine fracture detection system," *Electronics*, vol. 12, no. 9, p. 2056, Apr. 2023, doi: 10.3390/electronics12092056.

[16] Z. Krawczyk and J. Starzynski, "Bones detection in the pelvic area on the basis of YOLO neural network," in *Proc. 19th Int. Conf. Comput. Problems Electr. Eng.*, Sep. 2018, pp. 1–4, doi: 10.1109/CPEE.2018.8506970.

[17] (2022). *RSNA 2022 Cervical Spine Fracture Detection | Kaggle*. [Online]. Available: https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/data

[18] Y. Baba and A. Murphy, "Windowing (CT)," Radiopaedia.org, Tech. Rep., Mar. 2017. [Online]. Available: https://radiopaedia.org/articles/windowing-ct?lang=gb, doi: 10.53347/rID-52108.

[19] T. Dorosti, M. Schultheiss, F. Hofmann, J. Thalhammer, L. Kirchner, T. Urban, F. Pfeiffer, F. Schaff, T. Lasser, and D. Pfeiffer, "Optimizing convolutional neural networks for chronic obstructive pulmonary disease detection in clinical computed tomography imaging," 2023, *arXiv:2303.07189*.

[20] H. Lee, M. Kim, and S. Do, "Practical window setting optimization for medical image deep learning," 2018, *arXiv:1812.00572*.

[21] H. Ding, B. Wei, Z. Gu, Z. Yu, H. Zheng, B. Zheng, and J. Li, "KA-Ensemble: Towards imbalanced image classification ensembling under-sampling and over-sampling," *Multimedia Tools Appl.*, vol. 79, pp. 14871–14888, Jun. 2019, doi: 10.1007/s11042-019-07856-y.

[22] M. A. Arefeen, S. T. Nimi, and M. S. Rahman, "Neural network-based undersampling techniques," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 2, pp. 1111–1120, Feb. 2022, doi: 10.1109/TSMC.2020.3016283.

[23] N. I. Papandrianos, A. Feleki, S. Moustakidis, E. I. Papageorgiou, I. D. Apostolopoulos, and D. J. Apostolopoulos, "An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and grad-CAM," *Appl. Sci.*, vol. 12, no. 15, p. 7592, Jul. 2022, doi: 10.3390/app12157592.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019, doi: 10.1007/s11263-019-01228-7.

[25] M. Xiao, L. Zhang, W. Shi, J. Liu, W. He, and Z. Jiang, "A visualization method based on the grad-CAM for medical image segmentation model," in *Proc. Int. Conf. Electron. Inf. Eng. Comput. Sci. (EIECS)*, Changchun, China, Sep. 2021, pp. 242–247, doi: 10.1109/EIECS53707.2021.9587953.

[26] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[27] J. E. Small, P. Osler, A. B. Paul, and M. Kunst, "CT cervical spine fracture detection using a convolutional neural network," *Amer. J. Neuroradiol.*, vol. 42, no. 7, pp. 1341–1347, Apr. 2021, doi: 10.3174/ajnr.a7094.

[28] T. Meena and S. Roy, "Bone fracture detection using deep supervised learning from radiological images: A paradigm shift," *Diagnostics*, vol. 12, no. 10, p. 2420, Oct. 2022, doi: 10.3390/diagnostics12102420.

[29] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110190, doi: 10.1016/j.chaos.2020.110190.

[30] D. Zeyu, R. Yaakob, A. Azman, S. N. M. Rum, N. Zakaria, and A. S. A. Nazri, "A grad-CAM-based knowledge distillation method for the detection of tuberculosis," in *Proc. Int. Conf. Inf. Manag. (ICIM)*, Mar. 2023, pp. 72–77, doi: 10.1109/ICIM58774.2023.00019.

MUHAMMAD YASEEN received the B.E. degree in electrical engineering from Hamdard University, Pakistan. He is currently pursuing the master's degree with Inje University, Republic of Korea. His research interests include artificial intelligence, machine learning, deep learning, computer vision, and medical imaging.

MAISAM ALI received the B.E. degree in electrical and communication engineering from Hamdard University, Pakistan. He is currently pursuing the master's degree with Inje University, Republic of Korea. His research interests include artificial intelligence, machine learning, deep learning, and computer vision.

SIKANDAR ALI received the B.E. degree in computer engineering from the Mehran University of Engineering and Technology, Pakistan, and the M.S. degree from the Department of Computer Science, Chung Buk National University, Republic of Korea. He is currently pursuing the Ph.D. degree with the Department of Digital Anti-Aging Healthcare, Inje University, South Korea. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, computer vision, and medical imaging.

ALI HUSSAIN received the B.S.C.S. degree in computer science from the Government College University Faisalabad (GCUF), Pakistan, in 2019, and the master's degree from the Department of Digital Anti-Aging Healthcare, Inje University, South Korea, where he is currently pursuing the Ph.D. degree. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, computer vision, reinforcement learning, and medical imaging.

MOON-IL JOO received the Ph.D. degree in computer engineering from Inje University, South Korea, in 2018. He is currently a Research Professor with the Institute of Digital Anti-Aging Healthcare, Inje University. His research interests include software engineering, human–computer interaction, smart phone programming, and component-based development.

HEE-CHEOL KIM received the B.Sc. degree from the Department of Mathematics, the M.Sc. degree from the Department of Computer Science, Sogang University, South Korea, and the Ph.D. degree from the Numerical Analysis and Computing Science, Stockholm University, Sweden, in 2001. He is currently a Professor with the Department of Computer Engineering and the Head of the Institute of Digital Anti-aging Healthcare, Inje University, South Korea. His research interests include machine learning, deep learning, computer vision, and medical informatics.

• • •