Deep-Learning Classifiers for Small Data Orthopedic Radiology

Bilal Aslan, Winner Kazaka, Tomas Slaven,
Shaylin Chetty, Nicholas kruger, Geoff Nitschke
Department of Computer Science,
University of Cape Town, South Africa
aslbil001@myuct.ac.za, kzkwin001@myuct.ac.za, slvtom001@myuct.ac.za,
chtsha042@myuct.ac.za, nicholas.kruger@uct.ac.za, gnitschke@cs.uct.ac.za

***Abstract*—Training deep-learning classifiers in orthopedic pathology is problematic due to the scarceness of extensive datasets for training and testing meaning most orthopedic image data is small, sparse and noisy. This study evaluates the efficacy of various state-of-the-art supervised *Convolutional Neural Network* (CNN) image classifiers, complemented by data augmentation and transfer-learning, versus various *Neural Architecture Search* (NAS) based deep-learning classifiers. These classifiers are comparatively evaluated on two (cervical spine and elbow) small, multi-label (with unbalanced data distribution) orthopedic radiographic (X-ray) datasets, with the objective of detecting multiple pathologies with high accuracy. To bypass the pervasive problem of small datasets medical datasets, we implement pre-processing and layer freezing to boost all task performance metrics (accuracy, precision, recall, specificity, F1 score), with the *ResNet* CNN and *EfficientNet* classifiers yielding the best results overall. Results highlight the efficacy of applying specially tuned CNN and NAS classifiers to small, unbalanced and noisy datasets indicative of those used in orthopedic radiology, demonstrating the potential of such methods as automated prognostic and diagnostic tools to assist orthopedic practitioners.**

***Index Terms*—Neural Architecture Search, Deep-Learning, Multi-label Classification, Convolutional Neural Networks**

I. Introduction

Computer-Aided Diagnosis (CAD) systems have been widely adopted for medical imaging, lesion detection and diagnosis and even for prognosis prediction [1]. Most recently, deep-learning systems [2], have supported radiologists in data interpretation, help avoid human errors, and increase diagnostic accuracy [3]. For example, deep *Convolutional Neural Network* (CNN) classifiers have been trained for pathology classification in X-rays, assisting practitioner interpretations across various types of medical imaging including mammography [4], elbow joint effusion [5], chest [6] and musculo-skeletal [7] radiography. In radiography, prognosis prediction remains problematic given radiologist perception errors in image interpretation (interpretive errors [8]) resulting in missed diagnoses at an approximated 4% error rate [9]. Since one billion radiography examinations are performed worldwide annually this approximates to 40 million interpretive errors per year. Increasing patient numbers and unavoidable human perceptual limitations [10] mean that automated classifiers play an increasingly important role in decreasing interpretive error across diagnostic radiography.

978-1-7281-8393-0/21/$31.00 ©2025 IEEE

Recently, deep-learning CAD has been applied to pathology classification and prediction in cervical spine (neck) [11] and elbow [5] orthopedic radiology. The interpretation of such radiology data is usually not binary (normal or abnormal) [12], and while current deep-learning classifiers perform well for binary classification of a specific pathology per image, the prediction accuracy of multiple pathologies (multi-label classification) remains poor [13]. Also, most medical imaging data comprises large, high-resolution images (thus facilitating pathology identification), but many fewer images (average of 3500 images for bone fracture radiograph data [14]) compared to approximately one million images comprising datasets such as *ImageNet* [15]. Medical imaging also uses many fewer classes (average of 2-6 classes for bone fracture radiographs [14]) with often significant class imbalances, compared to, for example, 1000 balanced classes of *ImageNet* deep-learning [16]. So, while CAD tools assist some radiological interpretations, their real-world diagnostic accuracy is unclear [17] and critically dependent on suitable training data [18], which is not readily available in many cases.

Deep-learning advances for the purpose of automating medical imaging include multi-label image classification, where images have multiple labels and thus pathologies [19]. Multi-label classifiers, identifying multiple pathologies have been demonstrated across medical imaging tasks, predicting, for example, lung disease [20] and brain tumors [21]. However, applying multi-label deep-learning to orthopedic radiological datasets has received little research attention [13], and then only for specific imaging studies [11]. Few studies have however applied multi-label classifiers for multiple orthopedic pathology detection, type classification and localization in cervical spine and elbow radiographs [14].

Limited medical imaging training data availability remains a key challenge for deep learning, that can be addressed by *Data-Augmentation* (DA), a regularization technique suitable for mitigating over-fitting resulting from small, sparse and noisy data [22]. Several studies have proposed

DA methods suitable for medical imaging [23], reproducing data distributions close to real data [24], and improving deep-learning assisted disease diagnosis for various organs and imaging modalities (magnetic resonance, computed tomography and mammography) [25]. *Transfer-Learning* (TL) is a regularization strategy, where beneficial weight connections, CNN layers are transferred between source and target CNNs during training [26]. TL has been widely applied to medical imaging since it mitigates data scarcity and saves computational resources [23], [27]. A popular TL approach is to apply an established CNN architecture designed for natural image data (for example, *ImageNet*), using pre-trained weights (for example, *ResNet* [28]), and then fine-tune CNN performance on medical imaging data [16]. This approach has been used to train *ResNet*, *DenseNet* on chest X-rays [29]. Calibrating CNN architectures per training dataset demands expertise and time, motivating development of various *Neural Architecture Search* (NAS) methods such as DeepMAD [30], ZenNAS [31] and Co-Deep-NEAT [32]. NAS methods automatically adapt CNN topology and parameters yielding high image classification task-performance on established benchmarks (*ImageNet*, *CIFAR-100*) [33].

This study evaluates the efficacy of several state-of-the-art CNN and NAS classifiers for identifying a diverse range of pathology types, some of which are difficult for doctors to detect using only X-ray scans. The goal of this study is not to present completely new classify methods but rather to demonstrate that specific configurations (assemblies of established methods), including fine-tuning techniques such pre-processing and layer freezing, complemented by optimizers including DA and TL are an effective means for addressing the open problem of effective classification in orthopedic radiography image data classification (section II). That is, where such medical image data is small, sparse, and noisy and defined by multiple pathologies, necessitating multi-label classification.

TABLE I
CERVICAL SPINE (NECK) DATA: PATHOLOGIES.

Neck dataset			
Class Labels	**Train**	**Validation**	**Test**
Number of images	**746**	**93**	**94**
Alignment	411	51	52
Soft tissue swelling	83	10	10
Listhesis	63	11	8
Fracture	99	9	15
Dislocation	26	2	3
Spinous	43	6	3
Other pathogens	127	16	16
Normal	239	30	30

TABLE II
ELBOW DATA: PATHOLOGIES.

Elbow dataset			
Class Labels	**Train**	**Validation**	**Test**
Number of images	**2378**	**193**	**169**
Soft tissue swelling	202	31	31
Joint effusion	493	62	61
Distal humerus	64	12	9
Supracondylar	691	87	92
Medial epicondyle displaced	71	14	8
Lateral epicondyle displaced	111	12	18
Olecranon	52	6	6
Elbow dislocation anterior	12	2	2
Elbow dislocation posterior	47	2	2
Proximal radial	40	2	6
Radial head	14	2	8
Radial head subluxation	18	2	3
Proximal ulnar metaphysis	27	1	5
Normal	1255	57	17

II. METHODS AND EXPERIMENTS

Three CNN and NAS classifiers and a *Visual Transformer* (ViT) [34] were trained and evaluated on the neck and elbow datasets (Table I, II). The CNN classifier methods: *DenseNet* [35], *ResNet* [36], and *ConvNeXt* [37] and ViT classifier methods: *SwinTransformerv2* [34] were selected since they have demonstrated comparable average task performance for image classification, consistently across established computer vision bench-mark datasets including: *ImageNet-1K* and *ImageNet-22K*, as well as X-ray data [38]. Similarly, the NAS methods: *EfficientNetv2* [39], *DeepMAD* [30], and *ZenNAS* [31] were selected given demonstrated adapted architectures resulting in high (classification accuracy) task performance across bench-mark image datasets (CIFAR-10, ImageNet, NAS-Bench201, CIFAR-100), low training overhead (compute time), but high computational efficiency compared to related state-of-the-art methods [40]. All our method implementations [41] replicate those in previous work: *DenseNet* [35], *ResNet* [36], *SwinTransformerv2* [34], *EfficientNetv2* [39] and *ConvNeXt* [37], *DeepMAD* [30], and *ZenNAS* [31], and are thus not described here. All method parameters (table III) were specially tuned for given datasets in company with pre-processing, layer freezing, and transfer-learning and data-augmentation optimizers (section II). All other method parameters remained consistent with parameter settings tuned for previous work [30], [31], [34]–[37], [39].

Experiments[1] trained and evaluated 63 deep-learning classifiers (Table III) for multi-label (multiple pathology) classification given elbow and neck X-ray data (Tables II and I). Experiment set 1 applied CNN classifiers: *ConvNeXt*, *SwinTransformerv2*, *DenseNet* and *ResNet* and DA variants: *Random Cropping*, *Random Augment*, *Neural Augment* and TL (section II-B), to each dataset. Experiment set 2 applied NAS classifiers: *EfficientNetv2*, *DeepMAD*, and *ZenNAS*. All methods were implemented in Python 3.12 using the *PyTorch* framework. The models were pre-trained on the *ImageNet1K-V1* dataset to leverage transfer learning. All experiments were conducted on a high-performance computing (HPC) cluster equipped with Intel Xeon 24-core CPUs, 64GB of RAM, and Nvidia V100 GPUs for accelerated computation.

A. Evaluation Metrics

We selected classifier task-performance metrics specifically relevant to radiological imaging studies [42], including: *Average accuracy* and *Precision*, *Recall*, *Specificity* and *F1 score*, As per related work [43], equations 5-1 (TP: True Positive, FP: False Positive, FN: False Negative) present *F1 score* (using precision and recall metrics), and *accuracy* calculation.

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

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

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

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

$$F1score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{5}$$

B. Parameter Tuning and Data Processing

To enhance model generalization and improve the training process, we employed several techniques, including early stopping, which preserved the model iteration with the highest validation F1 score. Image normalization was conducted using the dataset-specific mean and standard deviation for each RGB channel. Additionally, data pre-processing involved converting image labels into one-hot encoded arrays, with binary values representing the presence or absence of each pathology class. The datasets were partitioned into training, validation, and testing subsets in an 80%, 10%, and 10% ratio, respectively.

C. Data Collection and Computation

Cervical spine and elbow radiograph image data [41], (Tables I, II) was sourced from an orthopedics department at a local hospital. In our dataset, we included images that exhibit abnormalities not frequently encountered in clinical practice. This dataset curation aims to investigate the capability of our methods to detect less common abnormalities, even in the absence of extensive representative data.

[1]All experiment results and classifier source code are available online: https://anonymous.4open.science/r/MedicalPathogenPredictionWithAI-E7D1

D. Transfer Learning (TL) and Data Augmentation (DA)

For TL experiments (experiment set 1), all methods were pre-trained on the *ImageNet1K-V1* dataset to leverage the generalization capabilities of large-scale image data. We evaluated these methods under three different fine-tuning scenarios: freezing all layers except the last two, freezing all layers except the last one, and unfreezing all layers for full network fine-tuning. These variations allowed us to assess the impact of layer-wise adjustments on model performance and determine the most effective strategy for adapting pre-trained models to our specific pathology detection task.

DA experiments applied *Random transformations* (including Random Cropping and Augment [44]) to all methods (experiment set 1). All images were resized to 224x224 then gray-scaled. For Random Augment, an image was first gray-scaled to nullify any color augmentations which would not maintain image integrity. Torch Vision's *Rand Aug* was applied with settings as per related work [39]. Figure 1 presents examples of these DA methods applied to a selected dataset image.

III. Results and Discussion

Table IV presents classification results for both datasets and all methods (section II), highlighting (in bold) the highest performing methods per dataset. Table IV presents the three and four best performing NAS and non-NAS classifiers, respectively, for the metrics: Accuracy, Precision, Recall, Specificity, and F1 score (section II-A). Pair-wise statistical tests (Mann–Whitney U, $p<0.05$, [45]) were conducted between all method result pairs (Table IV) with Effect Size [46] treatment. Results data [41] was non-parametric (KS normality test with *Lilliefors* correction [47]).

First, examining comparative average task performance across all metrics (section II-A) for non NAS methods (section II), applied to the elbow and neck datasets, we observe *ResNet* (*ResNet152*, *ResNet18*, Table IV) yielded significantly higher ($p<0.05$) F1 score while maintaining high specificity on the elbow and neck datasets, respectively, compared to the other non NAS classifiers. Second, examining task performance across all metrics, of the comparative NAS classifiers, we observe that *EfficientNetv2* yields the highest ($p<0.05$), F1 score while maintaining high specificity for both elbow and neck datasets, compared to the other NAS methods (Table IV).

These results are inline with related work [33], similarly demonstrating that NAS classifiers out-perform hand-crafted classifier architectures (including *DenseNet* [35] and *ConvNext* [37]), on various image classification datasets (including CIFAR10 and CIFAR100). For example, previous work [31] similarly demonstrated that EfficientNetv2 (highest performing NAS classifier, Table IV) out-performed other state-of-the-art deep-learning NAS classifiers, in terms of top-1 accuracy on ImageNet ILSVRC2012 [39]. These results also demonstrate the applicability and efficacy of EfficientNetv2 applied to small datasets with relatively few,

TABLE III
METHOD PARAMETERS (ARCHITECTURE AND HYPER-PARAMETERS). NAS: NEURAL ARCHITECTURE SEARCH.

Method Architectures	**Parameters (Layers, Approximate parameters)**
ConvNeXt-T / S / B	(12, 29x10^6) / (12, 50x10^6) / (12, 89x10^6)
Swinv2-T / S / B	(12, 28x10^6) / (24, 50x10^6) / (24, 88x10^6)
DenseNet-121 / 169 / 201	(121, 8x10^6) / (169, 14x10^6) / (201, 20x10^6)
Resnet-18 / 50 / 152	(18, 12x10^6) / (50, 26x10^6) / (152, 60x10^6)
EfficientNetv2-S / M / L (NAS)	([1, 15], 8x10^6 / 14x10^6 / 20x10^6)
DeepMAD (NAS)	(50, 24.2x10^6)
ZenNAS (NAS)	(50, 18.4x10^6)
Hyper-Parameter	**Value**
Cost Function	Binary Cross Entropy
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-4}/10 steps
Optimizer	Adam
Epochs	50
Batch Size	32
Image Size	224x224
Training Callbacks	Early Stopping, Normalization

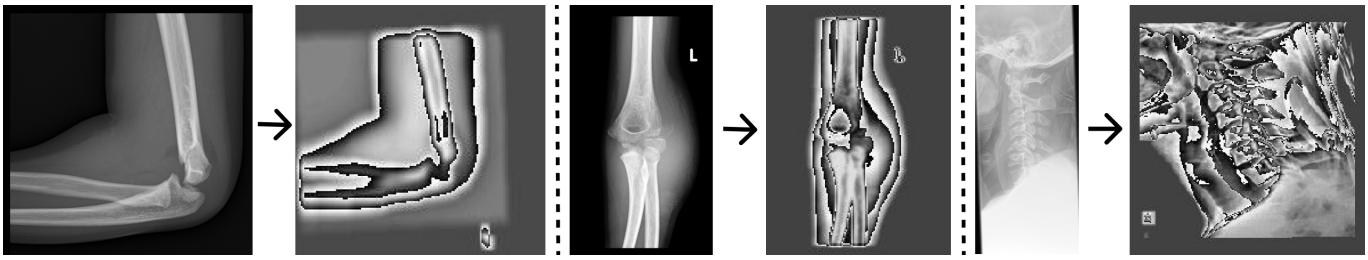

Fig. 1. **Left:** Lateral radiograph of the elbow showing a Gartland II supracondylar fracture, with the original image and corresponding result after DA application displayed to the left. **Center:** Anteroposterior (AP) radiograph of the elbow demonstrating soft tissue swelling, with the original and DA-processed images shown to the left. The 'L' marker in each image denotes the patient's left elbow. **Right:** Radiograph of the neck illustrating alignment, soft tissue swelling, listhesis, and fracture, with the original and DA-processed images on the left.

unbalanced images per class. That is, 746 images with eight classes in neck dataset (table I), compared to 1.2 million images for ImageNet-1k and 1000 classes (with approximately equal numbers of images assigned per class).

The demonstrated efficacy of EfficientNetv2 on relatively small, unbalanced X-ray datasets results from the architecture search space being sufficiently large to comprise multiple network topologies, which includes connectivity and weights common to various established classifiers, including *ResNet* [36] and *MobileNet* [48], where the efficacy of such classifiers has been demonstrated across various image classification tasks for various datasets. This sufficiently large and diverse architecture search space coupled with the EfficientNetv2 evolutionary search process enabled the derivation of a classifier architecture (Table III) suitable for producing the highest F1 score and specificity averaged together (Table IV).

Overall, these results support the notion that deep-learning classifiers, adapting network topology (including number of layers, connectivity and connection weights) via NAS, are suitable and effective methods when applied to classification tasks of relatively small and noisy datasets, indicative of those used in multi-pathology classification for orthopedic imaging studies. Here, our dataset consisted of 3124 (neck and elbow) X-ray images comprising 21 classes corresponding to present or absent pathologies (Tables II, I).

Related work has similarly applied deep-learning classifiers (also used here, Table III), including *DenseNet-201* and *ResNet-152*, to detect specific pathologies in elbow datasets [5], [49] (for example, joint effusion or fractures). These applications achieved comparable (86% [5] and 96% [49] average accuracy) given datasets ranging from 1032-4423 X-ray images. However, such work used larger datasets for training and testing, and classifiers were only tasked with the

TABLE IV

BEST METHOD TASK-PERFORMANCE ON NECK AND ELBOW TEST DATASETS FOR TASK-PERFORMANCE METRICS: AVERAGE ACCURACY, PRECISION, RECALL, SPECIFICITY, F1-SCORE. BEST NON-NAS AND NAS CLASSIFIERS HIGHLIGHTED IN RED. NAS: NEURAL ARCHITECTURE SEARCH. METHODS YIELDING OVERALL HIGHEST TASK-PERFORMANCE (ACROSS ALL METRICS) ARE HIGHLIGHTED IN BOLD.

Method Architecture	Accuracy	Precision	Recall	Specificity	F1 Score
Elbow Dataset					
ConvNeXtL (CNN)	89.35	97.18	90.79	76.47	93.88
ResNet152 (CNN)	**92.31**	**97.28**	**94.08**	**76.47**	**95.65**
DenseNet121 (CNN)	91.12	96.60	93.42	70.59	94.99
Swinv2t (ViT)	89.35	95.27	92.76	58.82	94.00
EfficientNetv2m (NAS)	**88.17**	**99.25**	**87.5**	**94.12**	**93.01**
DeepMAD (NAS)	55.03	90.43	55.92	47.06	69.10
ZenNAS (NAS)	65.09	92.66	66.45	52.94	77.39
Neck Dataset					
ConvNeXtL (CNN)	76.60	78.38	90.63	46.67	84.05
ResNet18 (CNN)	**81.91**	**87.30**	**85.93**	**73.33**	**86.61**
DenseNet161 (CNN)	78.72	80.56	90.62	53.33	85.29
Swinv2t (ViT)	74.47	75	93.75	33.33	83.33
EfficientNetv2S (NAS)	**80.85**	**81.08**	**93.75**	**53.33**	**86.96**
DeepMAD (NAS)	62.77	77.36	64.06	60.0	70.09
ZenNAS (NAS)	50.0	69.77	46.88	56.67	56.07

classifying one pathology. Whereas, in our study, classifiers were tasked with correctly classifying the presence or absence of 14 possible elbow pathologies (Table II).

A key contribution of this study is that, by comparison, there is relatively little related work classifying multiple pathologies given cervical spine datasets. For example, a recent review of automated image analysis studies using neck radiological imaging data [11], revealed most publications focused on the lumbar (*thoraco*) spine, with few deep-learning methods applied to cervical spine datasets. Specifically, of 19 spine imaging studies (2007-2020), only one applied deep-learning (*U-Net* architecture [50]) for pathology feature classification given X-ray data [51]. None of the reviewed studies applied classifiers to assist with pathology prediction (as demonstrated in this study).

Another key difference between this study and related work [52], [53] (similarly demonstrating the efficacy of established classifiers including *ConvNeXt-T*, *ResNet-152* and *Swinv2-T* applied to detect multiple pathologies given elbow and neck datasets), is that such related work invariably used image data generated from other modalities including magnetic resonance imaging, computer tomography and ultra-sound imaging devices. This is an important distinction, since compared to image data produced by such modalities, X-ray images are low quality and noisy and thus typically require pre-processing and image enhancements [54]. This adds to the complexity of automated multi-label (multi-pathology) classification using X-ray datasets, especially when such data are small with unbalanced distributions per class (indicative of orthopedic datasets used in this study, Table I, II).

Supporting the classification complexities inherent in this study (compared to related work [11], [52], [53]), is that the interpretation of pathogens from neck and elbow X-ray images presents significant challenges for both medical practitioners and deep-learning classifiers. For example, accurately predicting pathogens in neck X-ray images requires accurate observation and interpretation of the condition of all seven cervical vertebrae (C1 to C7). Identifying abnormalities is complicated due to the natural variability in vertebral alignment that can occur without pathological significance. Also, distinguishing between normal and abnormal variations in alignment and pathological conditions such as vertebral *listhesis* or dislocations poses additional challenges. These factors collectively contribute to the difficulty in predicting specific abnormalities from neck X-ray images. Accurate classification of elbow X-ray images is similarly complicated. For example, these images included two standard radiographic projections: *Lateral* and *Anteroposterior*, providing distinct anatomical views. The elbow dataset thus required our classifiers to differentiate between these orientations, while correctly classifying closely related abnormalities.

Furthermore, we observed that freezing certain layers of the pre-trained models improved performance, despite these models being initially trained on the ImageNet dataset,

which contains no X-ray images. This is an intriguing finding, as freezing layers helped mitigate the bias towards predictions—a common issue when training models on small datasets. Specifically, our results for elbow pathology classification (Table IV) revealed that *ResNet* exhibited the best performance across all metrics when only the last two layers were unfrozen, with the remaining layers frozen. This suggests a potential strategy for training deep learning models on medical datasets: pre-training on large, diverse image datasets followed by selective freezing of layers during fine-tuning on smaller, specialized medical datasets.

This study's key contribution was to demonstrate the benefit of applying specific (specially configured) NAS and non-NAS deep-learning classifiers for multi-pathology classification given small, sparse, unbalanced and noisy radiographic (X-ray) image data indicative of current orthopedic datasets. Specifically, this study demonstrated the necessary pre-processing and layer-freezing, transfer-learning and data-augmentation that must be used in company with training of these classifiers (ResNet and EfficientNetv2), if high F1 score with high specificity is to be yielded.

IV. Conclusion

This study applied various state-of-the-art *Convolutional Neural Network* (CNN) and *Visual Transformer* (ViT) classifiers versus various *Neural Architecture Search* (NAS) based classifiers to X-ray image orthopedic (neck, elbow) datasets, given a range of classification performance metrics. We demonstrated that completely novel ViT or NAS methods are not strictly necessary to address the problem of attaining high task performance classification on small, sparse, multi-label, unbalanced image datasets (indicative of orthopedic radiographic data). Rather, results demonstrate that established state-of-the art CNN and NAS methods, specially tuned with pre-processing and layer freezing and assembled together with specific optimizers such as transfer-learning and data-augmentation, are suitable for yielding very high F1 Score with high specificity (correctly detecting the presence or absence of multiple pathologies per X-ray image). Results indicate that the *ResNet* CNN and *EfficientNet* NAS classifiers yield the highest F1 Score with highest specificity overall, indicating suitable applicability of these classifiers to small, unbalanced, multi-pathology datasets in orthopedic radiology and thus potentially as computational prognostic and diagnostic tools to assist orthopedic specialists.

Future work will apply a broader range of NAS and non-NAS deep-learning classifiers and orthopedic datasets with the objective of devising automated computational tools (digital assistants) that complement the prognoses and diagnoses of radiologists and orthopedic surgeons, to help prevent diagnosis and prognosis errors [55]. An end goal is to automate the continued adaptation of such digital assistants in concert with increasing orthopedic (radiological) datasets, as part of larger research effort to produce perpetually adapting and self-sustaining autonomous systems [56].

References

[1] K. Doi, "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, p. 198–211, 2007.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] R. Aggarwal *et al.*, "Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-Analysis," *npj Digital Medicine*, vol. 4, no. 65, pp. doi.org/10.1038/s41 746–021–00 438–z, 2021.

[4] Y. Shen *et al.*, "Artificial Intelligence System Reduces False-positive Findings in the Interpretation of Breast Ultrasound Exams," *Nature Communications*, vol. 12, no. 5645, pp. doi.org/10.1038/s41 467–021–26 023–2, 2021.

[5] J. Huhtanen *et al.*, "Deep Learning Accurately Classifies Elbow Joint Effusion in Adult and Pediatric Radiographs," *Scientific Reports*, vol. 12, no. 11803, pp. doi.org/10.1038/s41 598–022–16 154–x, 2022.

[6] H. Shin *et al.*, "Diagnostic Performance of Artificial Intelligence Approved for Adults for the Interpretation of Pediatric Chest Radiographs," *Scientific Reports*, vol. 12, no. 10215, pp. doi.org/10.1038/s41 598–022–14 519–w, 2022.

[7] M. He *et al.*, "A Calibrated Deep Learning Ensemble for Abnormality Detection in Musculoskeletal Radiographs," *Scientific Reports*, vol. 11, no. 9097, pp. doi.org/10.1038/s41 598–021–88 578–w, 2021.

[8] M. Bruno, E. Walker, and H. Abujudeh, "Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction," *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[9] D. Sabih, "Image Perception and Interpretation of Abnormalities; Can we Believe our Eyes? Can we do Something about It?" *Insights Imaging*, vol. 2, no. 1, pp. 47–55, 2011.

[10] Y. Kim and L. Mansfield, "Fool me twice: Delayed Diagnoses in Radiology with Emphasis on Perpetuated Errors," *American Journal of Roentgenology*, vol. 202, no. 1, pp. 465–470, 2014.

[11] C. Goedmakers *et al.*, "Machine Learning for Image Analysis in the Cervical Spine: Systematic Review of the Available Models and Methods," *Brain and Spine*, vol. 2, no. 101666, 2022.

[12] A. Brady *et al.*, "Discrepancy and Error in Radiology: Concepts, Causes and Consequences," *Ulster Medical Journal*, vol. 81, no. 1, pp. 3–9, 2012.

[13] J. Lee and S. Chung, "Deep Learning for Orthopedic Disease Based on Medical Image Analysis: Present and Future," *Applied Sciences*, vol. 12, no. 681, p. doi.org/10.3390/app12020681, 2022.

[14] B. Hill *et al.*, "Deep Learning and Imaging for the Orthopaedic Surgeon: How Machines "Read" Radiographs," *Journal of Bone and Joint Surgery*, vol. 104, no. 18, pp. 1675–1686, 2023.

[15] J. Deng *et al.*, "ImageNet: A large-scale Hierarchical Image Database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA: IEEE, 2009, pp. 248–255.

[16] M. Raghu *et al.*, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Proceedings of the Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, 2019.

[17] L. Plesner *et al.*, "Commercially Available Chest Radiograph AI Tools for Detecting Airspace Disease, Pneumothorax, and Pleural Effusion," *Radiology*, vol. 308, no. 3, p. e231236, 2023.

[18] P. Ball, "Is AI Leading to a Reproducibility Crisis in Science?" *Nature*, vol. 624, no. 1, pp. doi.org/10.1038/d41 586–023–03 817–6, 2023.

[19] S. Anwar *et al.*, "Medical Image Analysis using Convolutional Neural Networks: A Review," *Journal of Medical Systems*, vol. 42, pp. 226–239, 2018.

[20] M. Al-Sheikh *et al.*, "Multi-class Deep Learning Architecture for Classifying Lung Diseases from Chest X-Ray and CT images," *Scientific Reports*, vol. 13, no. 19373, pp. doi.org/10.1038/s41 598–023–46 147–3, 2023.

[21] F. Yousaf *et al.*, "Multi-class Disease Detection using Deep Learning and Human Brain Medical Imaging," *Biomedical Signal Processing and Control*, vol. 85, no. 104875, 2023.

[22] G. Varoquaux and V. Cheplygina, "Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future," *npj Digital Medicine*, vol. 5, no. 48, pp. doi.org/10.1038/s41 746–022–00 592–y, 2022.

[23] C. Shorten and T. Khoshgoftaa, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 60, pp. 1–48, 2019.

[24] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review," *Journal of Imaging*, vol. 9, no. 4, p. doi.org/10.3390/jimaging9040081, 2023.

[25] E. Goceri, "Medical Image Data Augmentation: Techniques, Comparisons and Interpretations," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 12 561–12 605, 2023.

[26] K. Weiss, T. Khoshgoftaar, and D. Wang, "A Survey of Transfer Learning," *Journal of Big Data*, vol. 3, no. 9, pp. doi.org/10.1186/s40 537–016–0043–6.

[27] H. Kim *et al.*, "Transfer Learning for Medical Image Classification: A Literature Review," *BMC Medical Imaging*, vol. 22, no. 69, pp. doi.org/10.1186/s12 880–022–00 793–7, 2022.

[28] K. He *et al.*, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE, 2016, pp. 770–778.

[29] X. Wang *et al.*, "Hospital-scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE, 2017, pp. 3462–3471.

[30] X. Shen *et al.*, "DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network," *arXiv*, vol. 2303.02165, 2023.

[31] M. Lin *et al.*, "Zen-NAS: A Zero-shot NAS for High-performance Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Nashville, USA: IEEE, 2021, pp. 347–356.

[32] S. Acton *et al.*, "Efficiently Coevolving Deep Neural Networks and Data Augmentations," in *Proceedings of the IEEE Symposium Series on Computational Intelligence*. Canberra, Australia: IEEE, 2020, p. doi:10.1109/SSCI47803.2020.9308151.

[33] Y. Xu and Y. Ma, "Evolutionary Neural Architecture Search Combining Multi-branch ConvNet and Improved Transformer," *Scientific Reports*, vol. 13, no. 15791, pp. doi.org/10.1038/s41 598–023–42 931–3, 2023.

[34] Z. Liu *et al.*, "Swin Transformer v2: Scaling up Capacity and Resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. New Orleans, USA: IEEE, 2022, pp. 12 009–12 019.

[35] G. Huang *et al.*, "Densely Connected Convolutional Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE, 2017, pp. 4700–4708.

[36] I. Bello *et al.*, "Revisiting ResNets: Improved Training and Scaling Strategies," *arXiv*, vol. 2103.07579, 2021.

[37] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. New Orleans, USA: IEEE, 2022, pp. 11 976–11 986.

[38] T. Chauhan, H. Palivela, and S. Tiwari, "Optimization and Fine-tuning of DenseNet model for Classification of COVID-19 Cases in Medical Imaging," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. doi.org/10.1016/j.jjimei.2021.100020, 2021.

[39] M. Tan and Q. Le, "EfficientNetv2: Smaller Models and Faster Training," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[40] S. Khan *et al.*, "Transformers in Vision: A Survey," *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–41, 2022.

[41] Anon, "Anonymous Repository," *https://anonymous.4open.science/r/MedicalPathogenPredictionWithAI-E7D1*, 2024.

[42] S. Padash *et al.*, "An Overview of Machine Learning in Orthopedic Surgery: An Educational Paper," *Journal of Arthroplasty*, vol. 38, no. 10, pp. 1938–1942, 2023.

[43] B. Innocenti, Y. Radyul, and E. Bori, "The Use of Artificial Intelligence in Orthopedics: Applications and Limitations of Machine Learning in Diagnosis and Prediction," *Applied Sciences*, vol. 12, no. 10775, p. doi.org/10.3390/app122110775, 2022.

[44] L. Taylor and G. Nitschke, "Improving Deep Learning with Generic Data Augmentation," in *IEEE Symposium Series on Computational Intelligence*. IEEE, 2018, pp. 1542–1547.

[45] B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes*. Cambridge, UK: Cambridge University Press, 1986.

[46] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Milton Park, UK: Routledge, 1988.

[47] A. Ghasemi and S. Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non-statisticians," *International Journal of Endocrinology and Metabolism*, vol. 10, pp. 486–489, 2012.

[48] M. Sandler *et al.*, "MobileNetv2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE, 2018, p. doi:10.1109/CVPR.2018.00474.

[49] R. Jones *et al.*, "Assessment of a Deep-learning System for Fracture Detection in Musculoskeletal Radiographs," *npj Digital Medicine*, vol. 3, no. 144, pp. doi.org/10.1038/s41 746–020–00 352–w, 2020.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich, Germany: Springer, 2015, pp. 234–241.

[51] Y. Shin, K. Han, and Y. Lee, "Temporal Trends in Cervical Spine Curvature of South Korean Adults Assessed by Deep Learning System Segmentation, 2006-2018," *JAMA Network Open*, vol. 3, no. 10, p. e2020961, 2020.

[52] P. Chea and J. Mandell, "Current Applications and Future Directions of Deep Learning in Musculoskeletal Radiology," *Skeletal Radiology*, vol. doi.org/10.1007/s00256-019-03284-z, 2020.

[53] S.-J. Yoon *et al.*, "Automatic Multi-class Intertrochanteric Femur Fracture Detection from CT Images based on AO/OTA Classification using Faster R-CNN-BO Method," *Journal of Applied Biomedicine*, vol. 8, no. 4, pp. 97–105, 2020.

[54] I. Abedeen *et al.*, "FracAtlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs," *Scientific Data*, vol. 10, no. 521, pp. DOI:10.1038/s41 597–023–02 432–4, 2023.

[55] N. Regnard *et al.*, "Assessment of Performances of a Deep Learning Algorithm for the Detection of Limbs and Pelvic Fractures, Dislocations, Focal Bone Lesions and Elbow Effusions on Trauma X-rays," *European Journal of Radiology*, vol. 154, no. 110447, p. DOI:10.1016/j.ejrad.2022.110447, 2022.

[56] G. Nitschke and D. Howard, "AutoFac: The Perpetual Robot Machine," *IEEE Transactions on Artificial Intelligence*, vol. 3, pp. 2–10, 2022.