Proceedings of 15th IOE Graduate Conference
Peer Reviewed
Year: 2024 Month: May Volume: 15
ISSN: 2350-8914 (Online), 2350-8906 (Print)

A Hybrid Deep Learning Model for Musculoskeletal Abnormality Detection

Sudip Dahal [a], Ramesh Thapa [b], Santosh Panth [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pashchimanchal Campus, IOE, Tribhuvan University, Nepal*

[a] dahalsudip68@hotmail.com, [b] rthapa@wrc.edu.np, [c] santosh.panth51@gmail.com

Abstract

In the field of medicine, Deep Convolutional Neural Networks (DCNNs) have made significant strides, yet they face limitations in capturing comprehensive structural information due to their restricted perception capabilities.To address this limitation, this research proposes a novel approach that combines the strengths of VGG-16 for local detail extraction and Vision Transformer (ViT) for handling global features within images. With a focus on improving the classification of radiograph images, particularly in detecting Musculoskeletal Abnormalities, the study utilizes the MURA dataset consisting of 40,005 radiographic images. Through the integration of VGG-16 and ViT models, the research aims to achieve a more comprehensive analysis of radiograph images by capturing both local and global features effectively. The proposed methodology demonstrates promising results, with the hybrid model achieving an overall accuracy of 82.88% on the test set, along with a sensitivity of 0.8824.

Keywords

DCNNs, MURA, VGG-16, ViT

1. Introduction

In recent years, the demand for medical imaging, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, and X-rays, has surged dramatically. This surge in demand has outpaced the capacity of radiologists, particularly in low and middle-income countries. As a result, there is an urgent need to address the growing demand and availability of medical imaging for disease diagnosis [1].Manually processing medical data is frequently a time-intensive task, and the likelihood of misinterpretation errors cannot be ignored. Research suggests that in radiology, daily error rates and inconsistencies may exceed 3–5% [2].

Musculoskeletal disorders, which affect the bones, muscles, and joints, are a major cause of disability worldwide. They're not just a problem for older people but can affect anyone at any age. This puts a lot of pressure on radiologists, who often feel overwhelmed by their workload. To help them out, researchers are looking into using artificial intelligence (AI) to assist with diagnosing these disorders, especially in places like primary care clinics where there's a high demand for radiology services. AI can help analyze complex X-rays and other images to make diagnoses more accurate. With better technology and more data available, AI algorithms are getting better at this task, sometimes even outperforming humans [3].

In recent times, there has been a surge in research dedicated to identifying bone abnormalities. However, the majority of these studies narrow their focus to a single type of anomaly, which diminishes their applicability in real-world clinical scenarios. This limitation is justified by several factors, including the limited availability of public datasets, the diverse shapes of bones, and the wide spectrum of abnormality types. Consequently, crafting a dependable Computer-Aided Diagnosis (CAD) [4] system for bone abnormalities proves to be a formidable technical endeavor.

Recent advancements in deep learning have introduced more efficient methods for handling complex radiographic data.These innovations involve utilizing multiple hidden or fully connected layers[5] in training models, thereby enhancing their effectiveness. The Transformer is a new type of network that relies only on attention mechanisms, without using the usual recurrence and convolution methods. It includes an encoder and a decoder connected through attention.When tested on translation tasks, it showed better quality, could be run more tasks at the same time, and took less time to train compared to older models [6].

This study focuses on developing a hybrid deep learning model (VGG16+ViT) to predict musculoskeletal abnormalities using radiographic images.This innovative method showcases the potential of integrating diverse deep learning architectures, marking a significant step forward in addressing challenges related to musculoskeletal disorder diagnosis from radiographic images. A normal radiograph with no disease and an abnormal radiograph with diseases is shown in Fig 1.

Figure 1: Normal vs Abnormal Radiographs

2. Related Works

This section presents an overview of previous research conducted by various scholars. Rajpurkar et al.[7] previously utilized a 169-layer CNN for detecting upper extremity abnormalities in musculoskeletal images. However, their model yielded a relatively low accuracy of 38.9% when applied to finger radiographs. Chada G. [8] subsequently enhanced the results through deep transfer learning, while Verma M. et al.[9] concentrated on lower extremities using a densely connected CNN. Additionally, other researchers explored architectures such as VGG-19 and ResNet [10], achieving an accuracy of 82.13%, as well as utilizing Efficient-Net ensembles[11].

Ensemble learning and transfer learning with preprocessing were also employed, with a peak finger accuracy of 67.05% [12]. Gurpreet Singh [13] employed ComDNet-512 model, employing the Deflate compression technique and achieved highest accuracy of 89.41% in state-of-art method but model was trained with a limited dataset of around 4000. Badgeley et al. [14] propose using of an ensemble model that combines the outputs of a CNN with a classifier made with patient healthcare variables and achieved comparable results with the state-of-the-art.

El-Saadawy[15] utilized a two-stage approach combining GNG Network and VGG model for bone X-ray classification and abnormality detection, achieving the highest accuracy of 78.51%.Fang et al.[16] obtained an overall accuracy of 73.4% with the proposed iterative fusion CNN (IFCNN) method for classifying the MURA dataset. Pelka et al. achieved the highest accuracy of 79.85% using the InceptionV3 model on the entire MURA dataset [17]. Varma et al. achieved an AUC score of 0.88 in classifying the Lower Extremity Radiographs Dataset (LERA) by employing an ImageNet and DenseNet161 model pre-trained with MURA for the classification task[18].

Guan et al.[19] conducted studies indicating an average precision (AP) value of 62.04% for fracture detection. This was accomplished through the utilization of a deep CNN model, which involved marking fractures on arm X-ray images from the MURA dataset by physicians. In another study by Harini et al. [20], classification experiments were performed on finger, wrist, and shoulder images within the MURA dataset using five different CNN-based deep learning methods, resulting in a maximum accuracy of 56.30%.

3. Methodology

3.1 Dataset Preparation

3.1.1 Dataset

The dataset comprises more than 40,000 X-ray images covering seven distinct body regions, including elbows, fingers, forearms, hands, humerus, shoulders, and wrists. It encompasses a diverse array of abnormalities and conditions, such as fractures, dislocations, osteoarthritis, and various musculoskeletal disorders. Each X-ray image in the MURA dataset is labeled with binary annotations indicating whether it exhibits normal or abnormal characteristics. These annotations were meticulously performed by certified radiologists to ensure the accuracy and consistency of the dataset.

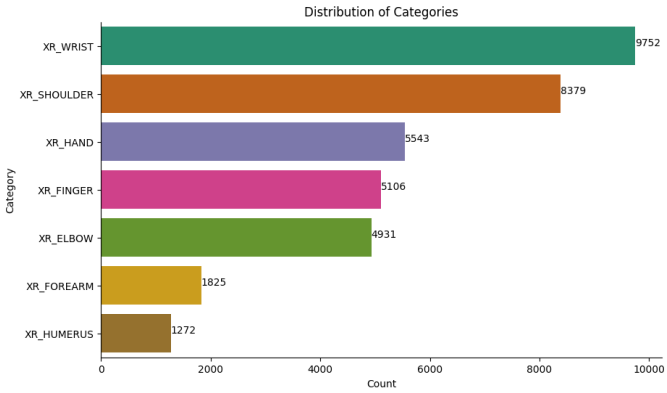

Figure 2: Distribution of dataset by categories in train-validation set

3.1.2 Dataset Preprocessing

The dataset collected was quite large. The image labels, including abnormal and normal from seven different classes, were stored in a CSV file and then imported into a dataframe using the pandas library. The file paths for each image were extracted and included in the dataframe. All labels were recognized, and one-hot encoding was applied. To ensure consistency, the images underwent normalization using the mean and standard deviation of the dataset. Following this, the images were resized to a uniform size of 224x224 pixels to best fit the data for classification.

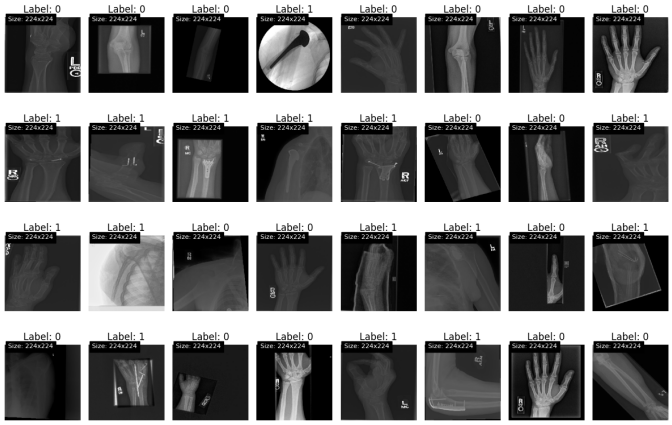

Figure 3: Visualization of dataset after preprocessing

3.1.3 Training and validation set

A total of 36,808 images were divided into training and validation sets using random shuffling. For the initial split, the proportion was 70:30, resulting in 25,766 images allocated to the training set and 11,042 images to the validation set. In the subsequent split, the proportion changed to 80:20, leading to 29,447 images in the training set and 7,361 images in the validation set.

Table 1: Number of images by category for train-validation set

Class	Normal	Abnormal
XR_ELBOW	2925	2006
XR_FINGER	3138	1968
XR_FOREARM	1164	661
XR_HAND	4059	1484
XR_HUMERUS	673	599
XR_SHOULDER	4211	4168
XR_WRIST	5765	3987
Total	**21935**	**14873**

3.1.4 Test set

The performance of the model was assessed using a distinct test dataset comprising 3197 images. These images had been evaluated and certified by board-certified radiologists from Stanford Hospital during clinical radiographic interpretation.

Table 2: Number of images by category for test set

Class	Normal	Abnormal
XR_ELBOW	235	230
XR_FINGER	214	247
XR_FOREARM	150	151
XR_HAND	271	189
XR_HUMERUS	148	140
XR_SHOULDER	285	278
XR_WRIST	364	295
Total	**1667**	**1530**

4. Model Architecture

4.1 Feature Extraction

4.1.1 VGG-16

VGG-16 is a convolutional neural network architecture designed for image classification tasks. It was proposed by the Visual Geometry Group (VGG) at the University of Oxford[21].In proposed architecture, the VGG-16 model is utilized as a feature extractor. It's employed without its fully connected layers, essentially as a pre-trained convolutional feature extractor for images. Once loaded, its layers are locked to retain the pre-trained weights and feature extraction abilities obtained from ImageNet. The output from VGG-16, after passing through its convolutional layers, is flattened to a one-dimensional vector.

4.1.2 ViT

ViT is a type of neural network architecture designed for computer vision tasks.The Vision Transformer model was proposed by Alexey Dosovitskiy et al. in the paper titled "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," published in 2020[22]. In this model setup, the Vision Transformer (ViT) serves as a complementary feature extractor alongside VGG-16. Unlike traditional convolutional networks, ViT captures global image patterns through self-attention mechanisms. Once loaded, ViT processes the input image, extracting high-level features that represent global relationships within the image.

4.2 Feature Concatenation

After extracting features from both VGG-16 and ViT models, they are merged together through concatenation. VGG-16 captures spatial features, while ViT focuses on global relationships. By combining these features, the model obtains a comprehensive representation containing both local and global information. Subsequent layers in the model refine and utilize these merged features for binary classification.

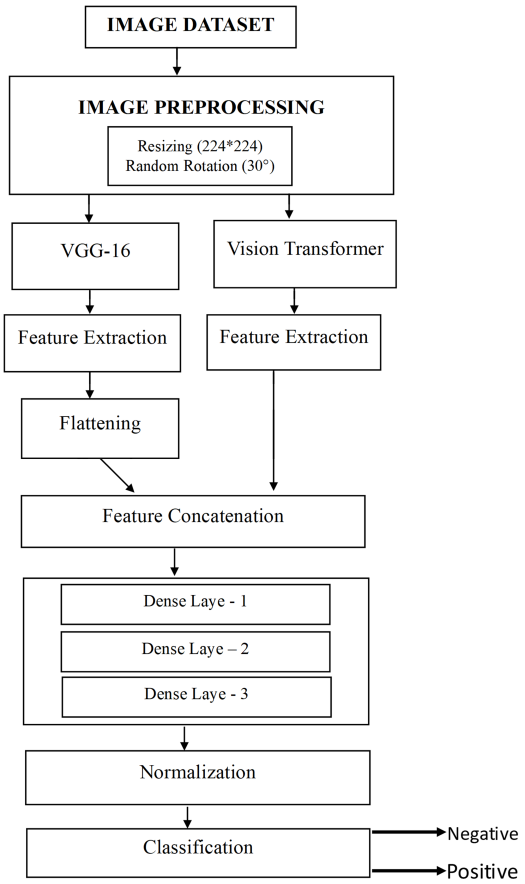

Figure 4: Block diagram of proposed model

4.3 Classification

A proposed model is constructed for binary classification by integrating features from both a pre-trained VGG16 model and a Vision Transformer (ViT) model. The input layer receives images of 224x224 pixels with RGB color channels. The VGG16 processes these images through convolutional and pooling layers, extracting hierarchical features. The Flatten layer converts the output of VGG16 into a 1-dimensional tensor for further processing. The ViT-B16, a Vision Transformer model, employs self-attention mechanisms to capture global information and relationships within the images.

The Concatenate layer combines the detailed features from VGG16 with the global context captured by ViT-B16. Dense layers further process these concatenated features, gradually reducing their dimensionality to learn intricate patterns. Finally, batch normalization stabilizes and accelerates training by normalizing outputs, enhancing overall efficiency and model performance.

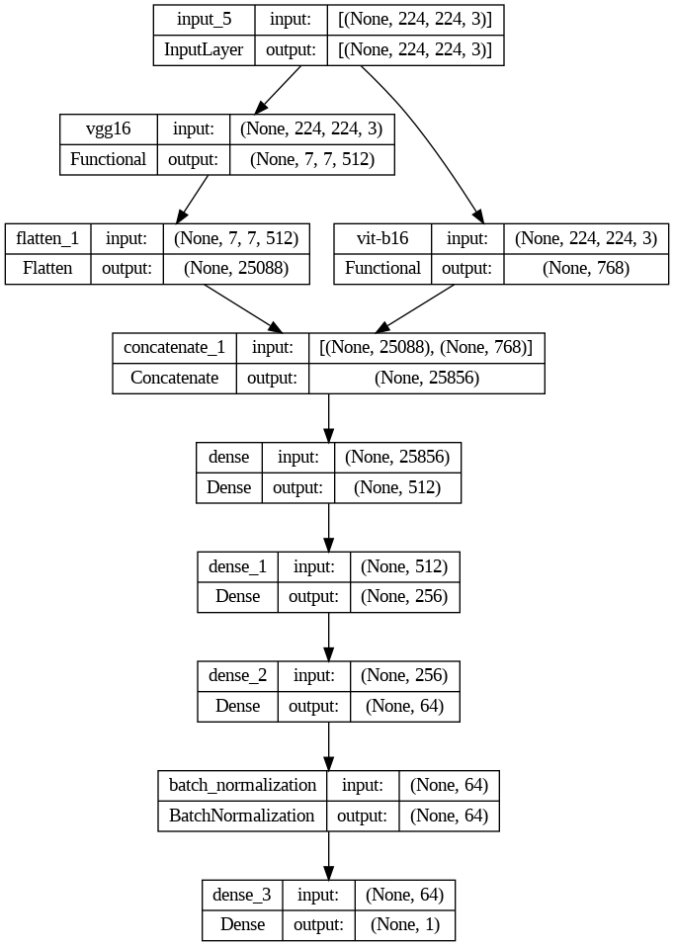

Figure 5: Model Architecture

5. Result and Discussion

5.1 Model training details

In the training process, we utilized pre-trained models, VGG16 and Vision Transformer (ViT), for feature extraction from image data. The layers of the VGG16 model were frozen to retain the valuable features learned from the ImageNet dataset. Our custom model architecture involved concatenating the output of the VGG16 and ViT models, followed by fully connected layers with GeLU activation functions and L2 regularization to mitigate overfitting. The model was compiled using the Adam optimizer with a various learning rate and binary cross-entropy as loss function. Evaluation metrics such as accuracy, precision, sensitivity and specificty were employed to assess the model's performance. During training, the model iterated over batches of data, minimizing the defined loss function through backpropagation. The validation dataset was used to monitor the model's generalization ability and prevent overfitting.

5.2 Results

The model's performance was evaluated using a distinct test dataset comprising 3197 images, all reviewed and validated by board-certified radiologists from Stanford Hospital during clinical radiographic interpretation. The following presents the outcomes obtained from different model assessments.Table 3 illustrates the results obtained from training the model using different hyperparameter configurations for ViT. The model's performance when trained with a batch size of 64, and utilizing a learning rate set at 0.0001. This combination notably led to the highest performance metrics for the pretrained ViT.

Table 3: Experiment Results on ViT using transfer learning

Model	Accuracy	Precision	Sensitivity	Specificity
ViT/B16	0.7648	0.7630	0.6457	0.8536

Table 4 depicts the outcomes derived from training the model under various hyperparameter setups for VGG-16. Among these, the model exhibited its best performance metrics when trained with a batch size of 64 and a learning rate of 0.00001.

Table 4: Experiment Results on VGG-16 using transfer learning

Model	Accuracy	Precision	Sensitivity	Specificity
VGG-16	0.7903	0.7825	0.6963	0.7864

Table 5 displays the results obtained from training the model under different hyperparameter configurations for VGG-16+ViT. The model achieved its highest performance metrics when trained with a batch size of 64, a learning rate set at 0.00001, and incorporating L2-Regularization with $\lambda = 0.0001$.

Table 5: Experiment Results on proposed VGG-16+ViT model

Epochs	Accuracy	Precision	Sensitivity	Specificity
30	82.88%	0.78	0.88	0.78

The loss curve and accuracy curve graphs illustrate the training process of a proposed model, showcasing fluctuations in performance metrics.

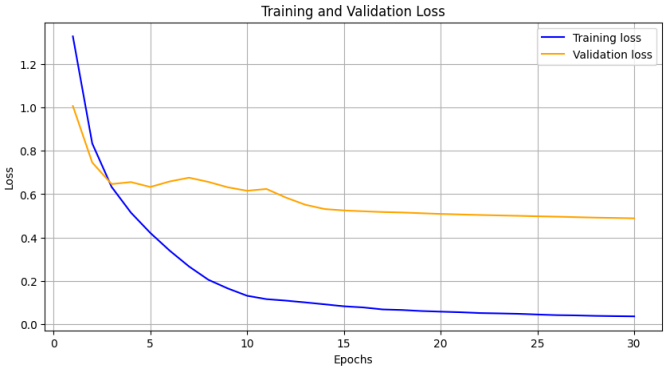

Figure 6: Training/Validation Loss Curve

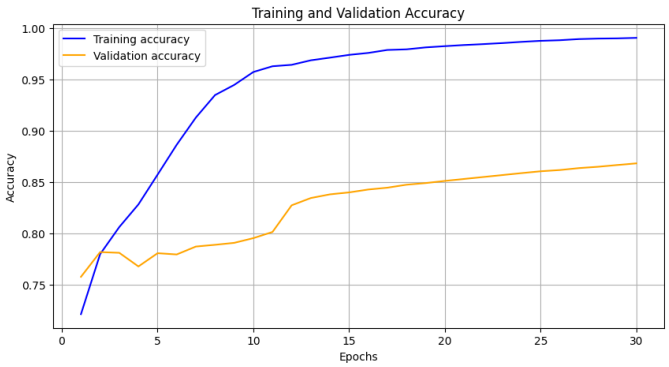

Figure 7: Training/Validation Accuracy Curve

The confusion matrix provided below sheds light on the model's performance in classifying musculoskeletal radiographs into normal and abnormal categories. Within the test set comprising 3197 images, the model accurately identified 1350 images as abnormal (True Positives), demonstrating its capability in detecting abnormalities effectively. However, there were 180 instances where abnormal images were mistakenly classified as normal (False Negatives), indicating areas of oversight. Conversely, the model correctly identified 1300 images as normal (True Negatives), highlighting its proficiency in recognizing radiographs without abnormalities. Nonetheless, there were 367 images falsely categorized as abnormal when they were normal (False Positives), suggesting instances of misclassification.

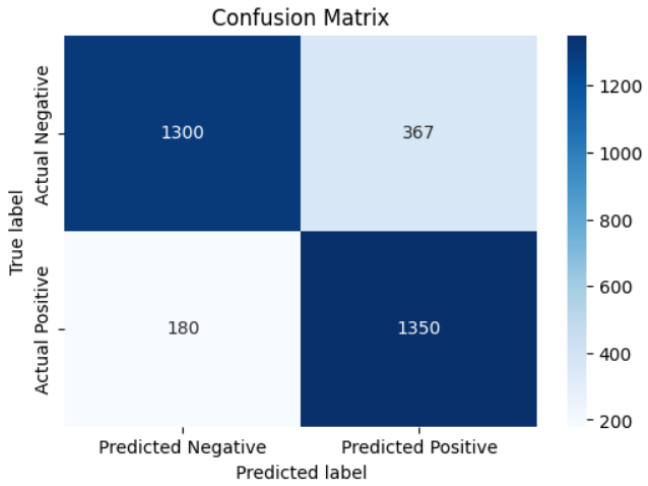

Figure 8: Confusion matrix

Table 6: Comparative performance analysis of VGG-16+ViT with state-of-art techniques

Model	Accuracy	Precision	Sensitivity
DenseNet-169	75.70%	0.88	0.86
DenseNet-201	76.57%	0.84	0.69
ConvNet	82%	0.86	0.72
VGG-16+ViT	**82.88%**	0.78	**0.88**

6. Conclusion

In this study, experiments were conducted on the MURA dataset to detect musculoskeletal abnormalities in upper extremities. Unlike prior research, which predominantly focused on binary classification within categories, the task was approached as a binary classification problem encompassing full dataset. This involved preprocessing the dataset and converting it into binary labels, encompassing images from all categories. The outcomes revealed promising results for the proposed VGG-16+ViT model, achieving an accuracy of 82.88%, precision of 0.7858, sensitivity of 0.8824, and specificity of 0.7804. These findings highlight the effectiveness of the approach in identifying musculoskeletal abnormalities in upper extremities. Sensitivity emerges as a crucial metric in medical image classification, reflecting the model's ability to accurately detect true positives—abnormalities present in the images. With a sensitivity of 0.8824, the model demonstrates strong capabilities for abnormality detection, emphasizing its importance in clinical diagnosis. It's noteworthy that no existing model has been trained on the complete dataset, underscoring the novelty and significance of this research. By prioritizing sensitivity, the study contributes to advancing automated abnormality detection, with potential implications for improving diagnostic accuracy and patient care in clinical settings.

However, this research has several limitations that should be acknowledged. Firstly, the performance of the model may be influenced by factors such as image quality, patient demographics, and variations in radiographic techniques, which were not explicitly accounted for in this study. Secondly, the interpretation of sensitivity and specificity values should consider the prevalence of abnormalities in the dataset, as they may be affected by class imbalances. Lastly, while the proposed VGG-16+ViT model shows promising results, further validation on independent datasets and clinical trials is necessary to assess its real-world utility and potential limitations in clinical practice.

Future work in this area could explore the integration of other deep learning architectures or ensemble methods to further enhance the model's performance. Additionally, incorporating domain-specific knowledge or expert annotations into the training process may improve the model's ability to detect subtle abnormalities. Moreover, conducting external validation studies on larger and more diverse datasets would help validate the generalizability of the proposed approach across different patient populations and imaging modalities.

References

[1] Govind Chada. Machine learning models for abnormality detection in musculoskeletal radiographs. *mdpi*, 2(4), Oct. 2019.

[2] Adrian P. Brady. An efficient deep neural network for disease detection in rice plant using xgboost ensemble learning framework. *Springer*, 8:171–182, Dec.07 2016.

[3] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human

performance? evidence from AI experts. *J. Artif. Intell. Res.*, 62:729–754, july 31 2018.

[4] H. El-Saadawy, M. Tantawi, H. A. Shedeed, and M. F. Tolba. A hybrid two-stage gng–modified vgg method for bone x-rays. *IEEE*, 9:76649 – 76661, May 29 2021.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, Honolulu, HI, USA, July 21–26 2017.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "attention is all you need". 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[7] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv*, 1712.06957v4, 2018.

[8] Govind Chada. Machine learning models for abnormality detection in musculoskeletal radiographs. *MDPI*, Oct.22 2019.

[9] M. Varma, M. Lu, R. Gardner, J. Dunnmon, N. Khandwala, P. Rajpurkar, J. Long, C. Beaulieu, K. Shpanskaya, L. Fei-Fei, et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nat. Mach. Intell.*, 1:578–583, 2019.

[10] T. C. Mondol, H. Iqbal, and M. Hashem. Deep cnn-based ensemble cadx model for musculoskeletal abnormality detection from radiographs. In *Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering*, pages 392–397, Dhaka, Bangladesh, 2019. IEEE, IEEE.

[11] K. Teeyapan. Abnormality detection in musculoskeletal radiographs using efficientnets. In *Proceedings of the 24th International Computer Science and Engineering Conference*, pages 1–6, Bangkok, Thailand, dec 2020.

[12] M. He, X. Wang, and Y. Zhao. A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. *Sci. Rep.*, 11:1–11, 2021.

[13] Gurpreet Singh, Darpan Anand, Woong Cho, Gyanendra Prasad Joshi, and Kwang Chul Son. Hybrid deep learning approach for automatic detection in musculoskeletal radiographs. *Biology*, 11(5):665, Feb. 18 2022.

[14] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(31), 2019.

[15] H. El-Saadawy, M. Tantawi, H. A. Shedeed, and M. F. Tolba. A hybrid two-stage gng–modified vgg method for bone x-rays classification and abnormality detection. *IEEE Access*, 9:76649–76661, 2021.

[16] L. Fang, Y. Jin, L. Huang, S. Guo, G. Zhao, and X. Chen. Iterative fusion convolutional neural networks for classification of optical coherence tomography images. *J. Vis. Commun. Image Represent.*, 59:327–333, 2019.

[17] O. Pelka, F. Nensa, and C.M. Friedrich. Branding-fusion of meta data and musculoskeletal radiographs for multi-modal diagnostic recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, Seoul, Korea, October 27–28 2019.

[18] M. Arma, M. Lu, R. Gardner, J. Dunnmon, N. Khandwala, P. Rajpurkar, J. Long, C. Beaulieu, K. Shpanskaya, L. Fei-Fei, and et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nat. Mach. Intell.*, 1:578–583, 2019.

[19] B. Guan, G. Zhang, J. Yao, X. Wang, and M. Wang. Arm fracture detection in x-rays based on improved deep convolutional neural network. *Comput. Electr. Eng.*, 81:1–11, 2020.

[20] N. Harini, B. Ramji, S. Sriram, V. Sowmya, and K.P. Soman. *Musculoskeletal Radiographs Classification using Deep Learning*. Academic Press, London, UK, 1st edition, 2020.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Visual Geometry Group, Department of Engineering Science, University of Oxford*, 2015.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. Google Research, Brain Team, 2021.