AI-Assisted Shoulder X-ray Classification for Enhancing Rehabilitation Decisions

Hana Hesham
Systems and Biomedical Engineering
Cairo University
hana.morsi02@eng-st.cu.edu.eg

Khaled Mohamed
Systems and Biomedical Engineering
Cairo University
khaled.abdelmawgood02@eng-st.cu.edu.eg

Omar Emad
Systems and Biomedical Engineering
Cairo University
omar.abdelbar02@eng-st.cu.edu.eg

Omar Nabil
Systems and Biomedical Engineering
Cairo University
omar.hussein03@eng-st.cu.edu.eg

Shehab Mohamed
Systems and Biomedical Engineering
Cairo University
shehabhegab20@gmail.com

***Abstract*—This study presents an AI-assisted system designed to classify shoulder X-ray images according to the Walsh classification of glenoid morphology, aiming to support rehabilitation professionals in making prosthetic decisions. The system utilizes a convolutional neural network based on the EfficientNetB1 architecture, fine-tuned on a curated dataset of 841 anterior-posterior shoulder X-ray images divided into four categories: A1, C1, D1, and Others. To address class imbalance, a combination of downsampling and data augmentation techniques was applied. The model achieved an overall classification accuracy of 86% on the test set, with strong performance particularly in identifying D1 and Other classes. The trained model was integrated into a desktop application to enhance accessibility and usability for clinical settings. This approach demonstrates the potential of deep learning in augmenting diagnostic and rehabilitation workflows for shoulder-related conditions.**

I. Introduction

Shoulder pathologies, particularly glenoid deformities, are common among aging populations and can significantly impact upper limb mobility and quality of life. Accurate classification of glenoid morphology is crucial in planning effective rehabilitation strategies and determining the need for prosthetic interventions. The Walsh classification system provides a standardized framework for identifying glenoid deformities such as backward-tilted (C1), forward-tilted (D1), and mildly worn (A1) glenoids, each with specific implications for prosthetic fitting and rehabilitation planning.

Traditionally, the evaluation of shoulder X-rays relies heavily on the expertise of radiologists. However, inter-observer variability and time constraints can affect the consistency and speed of diagnosis, especially in resource-limited or high-volume settings. In recent years, artificial intelligence (AI) and deep learning techniques have demonstrated promising results in medical image classification tasks, offering the potential to enhance clinical decision-making and support healthcare providers.

This project introduces a deep learning-based model for automated classification of shoulder X-ray images according to the Walsh classification. The model was trained on a labeled dataset of shoulder radiographs and deployed within a user-friendly desktop application designed to assist clinicians in rehabilitation environments. By integrating AI into routine diagnostic workflows, this system aims to improve diagnostic accuracy, reduce assessment time, and ultimately support better rehabilitation outcomes for patients with shoulder joint disorders.

II. Literature review

The application of deep learning in medical imaging has significantly advanced diagnostic capabilities, particularly in musculoskeletal radiology. Convolutional Neural Networks (CNNs) have been instrumental in automating the analysis of shoulder radiographs, facilitating the detection and classification of various pathologies.

Uysal et al. (2021) explored the classification of shoulder X-ray images using deep learning ensemble models. They evaluated multiple CNN architectures, including ResNet, DenseNet, and VGG, on the MURA dataset to distinguish between normal and abnormal shoulder images. Their ensemble approach achieved a test accuracy of approximately 84.7%, demonstrating the potential of deep learning models in shoulder pathology classification.

In another study, researchers developed a deep learning model to automate the measurement of the shoulder critical angle (CSA) and acromion index (AI) on anteroposterior shoulder radiographs. Utilizing the MURA dataset, the model achieved a mean absolute error of 1.68° for CSA and 0.03 for AI, indicating high precision in morphometric assessments.

Furthermore, a CNN-based approach was employed to detect and classify rotator cuff tears (RCTs) on plain shoulder radiographs. The model demonstrated a sensitivity of 92%, specificity of 69%, and an overall accuracy of 86% in detecting RCTs, outperforming orthopedic surgeons in diagnostic performance.

Additionally, Satir et al. (2024) developed an open-source deep learning model for the automatic quantification of scapular and glenoid morphology using CT scans. The model accurately measured parameters such as glenoid version and inclination, with results comparable to manual assessments, thereby aiding in the evaluation of glenohumeral osteoarthritis.

These studies underscore the efficacy of deep learning models in analyzing shoulder imaging, enhancing diagnostic accuracy, and supporting clinical decision-making. However, the application of AI specifically for classifying glenoid morphology according to the Walsh classification remains underexplored, highlighting the need for further research in this domain.

III. Dataset and Processing Steps:

Dataset:

The dataset used for training and evaluating the AI model consists of 841 anterior-posterior shoulder X-ray images, organized into three subsets: training, validation, and testing. Each subset contains subfolders corresponding to four classification categories based on the Walsh system: Type A1, C1, D1, and Others.

These images were retrospectively collected from pediatric patients aged 40 to 70 years at Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China. All imaging was conducted as part of routine clinical care.

Prior to training, the dataset underwent a thorough quality control process. All chest radiographs were initially screened to exclude low-quality or unreadable scans. The classification labels for each image were assigned by two expert physicians. To ensure annotation reliability, the evaluation set was independently reviewed by a third senior expert to correct for potential grading discrepancies.

The clinical relevance of the Walsh classes is as follows:

- A1: Normal glenoid with mild wear, typically the most straightforward case for prosthetic fitting.
- C1: Glenoid with posterior tilt, often requiring augmented implants or bone grafting.
- D1: Glenoid with anterior tilt, which may necessitate reverse shoulder arthroplasty.
- Others: Cases that do not clearly fit into the above classifications.

Processing Steps:

The dataset comprised 841 anterior-posterior shoulder X-ray images, categorized into four Walsh classification types: A1, C1, D1, and Others. The dataset was pre-divided into training (841 images), validation (104 images), and test (104 images) folders. Initial class distribution in the training set was significantly imbalanced, with the majority of samples labeled as Others (552 images), and only 57–172 images across the clinically relevant A1, C1, and D1 categories.

To address this imbalance and ensure fair learning, we applied a two-stage preprocessing approach: downsampling the majority class and augmenting underrepresented classes.

Downsampling the Majority Class

To create a balanced foundation, we downsampled the Others class to match the number of samples in the largest minority class (D1, with 172 samples). This resulted in a preliminary balanced training set of 461 images:

- A1: 57
- C1: 60
- D1: 172
- Others: 172

Data Augmentation for Minority Classes

To equalize all classes at 172 samples, data augmentation was applied to the underrepresented classes (A1 and C1). We applied random horizontal/vertical flips, rotations, width/height shifts, and zoom to synthetically increase sample diversity. The augmentation was conducted iteratively per class until each reached 172 samples.

After preprocessing, the final balanced training set included 688 images with an equal class distribution:

- A1: 172
- C1: 172
- D1: 172
- Others: 172

This preprocessing pipeline helped mitigate class imbalance and improve model generalizability across all glenoid types.

Figure 1 illustrates the outcome of our preprocessing pipeline across the X-Ray modality for some cases.

IV. Methods

Architecture

To classify shoulder X-ray images according to the Walsh classification system (A1, C1, D1, Others), we developed a convolutional neural network based on the EfficientNetB1 architecture. EfficientNet models are known for their high performance and computational efficiency, making them suitable for medical imaging tasks. The base model was initialized with ImageNet pretrained weights, excluding the top classification layers, and a global max pooling layer was used to reduce feature dimensions.

We added a batch normalization layer to stabilize training, followed by a dense layer with 256 units and ReLU activation. L2 kernel regularization ($\lambda = 0.016$), along with L1 activity and bias regularization ($\lambda = 0.006$), was applied to reduce overfitting. A dropout layer (rate = 0.45) was included to further improve generalization. The output layer consisted of a softmax activation function with four units corresponding to the Walsh categories.

The model was compiled using the categorical cross-entropy loss function and optimized with the Adamax optimizer (learning rate = 0.001). Accuracy was used as the evaluation metric.

To optimize training, we implemented a dynamic learning rate adjustment callback that reduced the learning rate when training accuracy plateaued (below 90%) or validation loss stopped improving. Early stopping was also applied, halting training after three stagnant epochs. The final model was trained for 40 epochs using data generators with batch-wise augmentation and validation support.

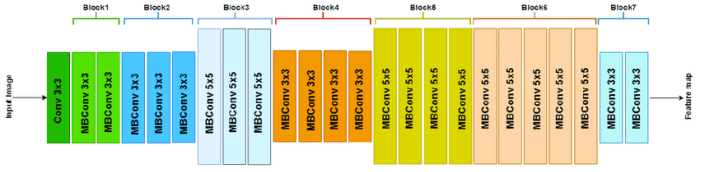

Figure 2: Architectural overview of the proposed EfficientnetB1. (source)

Loss Function And Optimization

The network was trained using categorical cross-entropy loss, defined as

$$L_{Categorical\ Cross\ Entropy} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\, log(P(c \mid x_i))$$

where N represents the number of samples, C denotes the number of classes, $y_{i,c}$ is the ground truth, and $P(c \mid x_i)$ represents the predicted probability.

Metrics

- **Precision**:
  Precision measures the ratio of correctly predicted positive voxels to all predicted positive voxels:
  - $Precision = \frac{TP}{TP + FP}$
- **Recall**:
  Recall (Sensitivity) measures the model's ability to identify true positives:
  - $Recall = \frac{TP}{TP + FN}$
- **F1-Score**:
  The harmonic mean of precision and recall, offering a balanced measure, especially useful for imbalanced datasets:
  - $F1.Score = 2 . \frac{Precision . Recall}{Precision + Recall}$

V. Experiments and results

The model was evaluated on an independent test set of 104 shoulder X-ray images distributed across the four Walsh categories. The performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score.
The trained EfficientNetB1 model achieved an overall accuracy of 86%. The class-wise performance metrics are shown in Table 1. Notably, the model performed best in detecting the "Others" and "D1" classes, achieving F1-scores of 0.92 and 0.79, respectively. However, the performance was lower for the underrepresented classes A1 and C1, likely due to limited sample sizes despite augmentation.

TABLE I

	Metric			
Subset	***Precision***	***Recall***	***F1-Score***	***Support***
A1	**1.00**	**0.57**	**0.73**	**7**
C1	**0.50**	**0.71**	**0.59**	**7**
D1	**0.68**	**0.95**	**0.79**	**20**
Others	**0.98**	**0.87**	**0.92**	**70**

The macro-averaged F1-score was 0.76, and the weighted average F1-score was 0.86, indicating strong overall performance despite the class imbalance. The confusion matrix further highlighted misclassifications between A1 and C1, which are clinically adjacent in diagnosis, emphasizing the need for additional data or domain-specific augmentation strategies.
Training and validation curves demonstrated stable convergence with minimal overfitting, aided by regularization and dropout. These results affirm the model's suitability for integration into clinical decision-making support tools in shoulder rehabilitation contexts.

Qualitative Results and Visualization

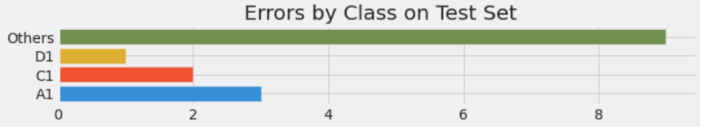

Figure 5 Displays Errors by Class on the Test set.

VI. Conclusion

In this project, we developed an AI-assisted shoulder X-ray classification system aimed at enhancing clinical decision-making in rehabilitation settings. Using a deep learning model based on the EfficientNetB1 architecture, our system successfully classified glenoid types according to the Walsh classification with an overall accuracy of 86%. To address class imbalance in the dataset, we applied a combination of downsampling and targeted data augmentation, which contributed to improved model performance across all categories.
Beyond model development, we deployed the AI system in a desktop application to ensure ease of use for healthcare professionals. This integration allows for quick and accessible analysis of shoulder X-rays, supporting more informed decisions regarding prosthetic needs and rehabilitation strategies.
Our results highlight the potential of deep learning in assisting clinical workflows, particularly in orthopedic and rehabilitation contexts. Future work will focus on expanding the dataset, exploring more advanced model architectures, and incorporating real-time clinical feedback to further validate and refine the system's utility in practical environments.

VII. Future Work

While the proposed system has shown promising results in classifying shoulder X-rays using the Walsh classification, several directions remain for future improvement:

Dataset Expansion:

The current dataset is limited in size, particularly for certain glenoid types. Acquiring a larger and more diverse dataset—potentially from multiple institutions—would improve the model's generalizability and robustness.

Clinical Validation:

Prospective validation in real clinical environments is essential. Collaborating with orthopedic clinics and rehabilitation centers will allow us to test the system's practical effectiveness and gather expert feedback.

Model Optimization:

Future versions could explore newer model architectures or ensemble approaches to boost classification accuracy, especially for underrepresented classes like A1 and C1.

Explainability and Trust:

Incorporating explainable AI (XAI) techniques, such as Grad-CAM, can help clinicians better understand the model's decisions and increase trust in its outputs.

VIII. References

[1] Uysal, E., Erbay, H., & Tutar, M. S. (2021). Classification of musculoskeletal radiographs using ensemble learning and deep convolutional neural networks. Applied Sciences, 11(6), 2723. https://doi.org/10.3390/app11062723

[2] Cheng, C. K., Wong, T. T., Chan, C. W., Ng, K. H., & Cheung, J. P. (2023). Artificial intelligence for automated measurement of shoulder critical angle and acromion index on radiographs. Journal of Shoulder and Elbow Surgery. https://pubmed.ncbi.nlm.nih.gov/37588867/

[3] Sato, R., Yamamoto, N., Tanaka, M., & Itoi, E. (2024). Deep learning model to detect rotator cuff tears on shoulder radiographs: A diagnostic performance study. Journal of Shoulder and Elbow Surgery. https://pubmed.ncbi.nlm.nih.gov/38311106/

[4] Satir, A., et al. (2024). GlenoNet: An open-source deep learning model for automatic quantification of scapular and glenoid morphology using CT. EPFL Research Repository. https://infoscience.epfl.ch/record/306038

[5] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML). https://arxiv.org/abs/1905.11946

[6] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint. https://arxiv.org/abs/1412.6980

[7] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1251–1258.

IX. Acknowledgment

We would like to express our heartfelt gratitude to Dr. Aliaa Rehan Youssef for her invaluable guidance. Her expertise was instrumental in shaping our work. We also extend our sincere appreciation to the teaching assistant, Ola Sarhan, for her dedicated assistance

X. Appendix

Appendix A: Training Performance Analysis

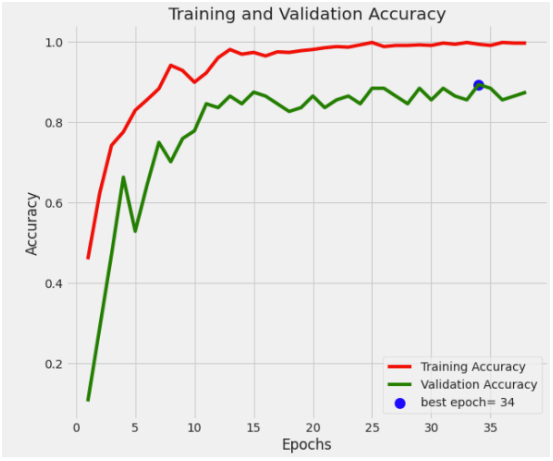

Figure A1: Training and Validation Overall Accuracy Progression for EfficientNetB1

Appendix B: Confusion Matrix of Test Set Classification Results

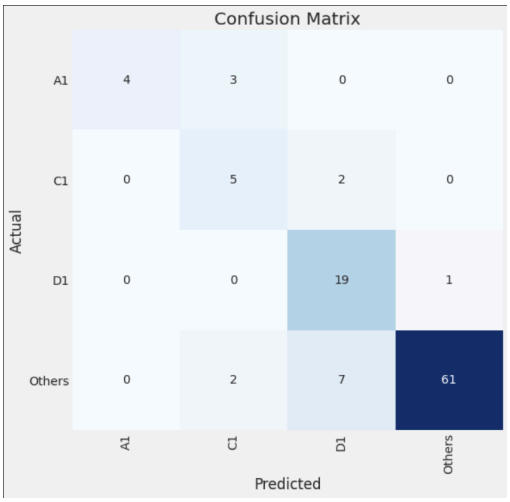

Figure B1: Confusion matrix showing model performance on the test set

.