Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

XGA-Osteo: Towards XAI-Enabled Knee Osteoarthritis Diagnosis with Adversarial Learning

Hieu Phan[1] , **Loc Le**[1] , **Mao Nguyen**[1] , **Phat Nguyen**[1] , **Sang Nguyen**[1] , **Minh-Triet Tran**[2] and **Tho Quan**[1*]
[1]Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam
[2]University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
{hieupt, loc.le, nxmao.sdh232, phat.nguyencs20, sang.nguyenvinh, qttho}@hcmut.edu.vn,
tmtriet@fit.hcmus.edu.vn

Abstract

This research introduces XGA-Osteo, an innovative approach that leverages Explainable Artificial Intelligence (XAI) to enhance the accuracy and interpretability of knee osteoarthritis diagnosis. Recent studies have utilized AI approaches to automate the diagnosis using knee joint X-ray images. However, these studies have primarily focused on predicting the severity of osteoarthritis without providing additional information to assist doctors in their diagnoses. In addition to accurately diagnosing the severity of the condition, XGA-Osteo generates an anomaly map, produced from a reconstructed image of a healthy knee using adversarial learning. Thus, the abnormal regions in X-ray images can be highlighted, offering valuable supplementary information to medical experts during the diagnosis process.

1 Introduction

Knee Osteoarthritis (KOA) is one of the most common types of degeneration. This disorder occurs when the safeguarding cartilage that covers the joints deteriorates, causing the bones to grind against each other devoid of the cushioning effect of the cartilage [Lespasio *et al.*, 2017]. The diagnosis process for this disease still heavily relies on manual assessment by doctors, which may lead to mistakes and inconsistencies due to their subjectivity and limitations. With the rise of artificial intelligence, numerous research groups have applied AI for automated disease diagnosis [Wahyuningrum *et al.*, 2016; Gu *et al.*, 2022; Swiecicki *et al.*, 2021]. Yet, all of these methods, provide only a single diagnosis, lacking explanatory capability to aid doctor's evaluation. Doctors, however, not only diagnose but also provide explanations about the patient's condition based on abnormal signs detected from knee X-ray images. Unfortunately, there is currently no labeled dataset available for segmenting these abnormal signs on X-ray images. Instead, anomaly detection techniques are used to identify these abnormalities [Georgescu, 2023; Iqbal *et al.*, 2024]. By integrating anomaly detection techniques, we can provide an anomaly map that assists doctors in the diagnostic process, thereby enhancing the system's effectiveness. We introduce the XGA-Osteo[1] , a specialized application designed to diagnose the severity of knee osteoarthritis. This innovative tool not only provides a diagnosis but also generates an anomaly map, highlighting at-risk areas on the patient's knee X-ray image.

Our main contributions are 3-fold. First, we introduce an AI framework called Osteo-GAN that leverages adversarial learning to reconstruct the healthy counterpart of a diseased knee X-ray image. Next, we introduce a method to generate an anomaly map from the reconstructed image, based on which one can identify abnormal signs in knee X-ray images. This allows us to identify these abnormalities without relying on labels, especially in the absence of any labeled datasets for segmenting these degenerated regions. Finally, we have developed an application called XGA-Osteo that offers knee osteoarthritis severity diagnosis in an explainable manner. By providing both information about abnormal regions in the patient's X-ray images and the severity of the disease, we expect this application to reduce the time required for the diagnosis process and provide valuable insights for doctors.

2 Related Work

2.1 Knee Osteoarthritis Diagnosis

In the research conducted by [Tiulpin *et al.*, 2018], the lateral and medial views of knee joint images were extracted to form pairs of inputs for the Siamese network to assess the disease severity. Another study used a Multi-Input CNN, as employed by [Swiecicki *et al.*, 2021], to combine two perspectives, PA and LAT, and improve classification performance. [Jain *et al.*, 2023] applied the *Convolutional Block Attention Module* (CBAM) to the feature map of the HRNet [Wang *et al.*, 2020] and achieved promising results. Self-attention mechanisms have also been effectively applied in computer vision, surpassing CNN performance when sufficient data is available. [Alshareef *et al.*, 2022] used Vision Transformer (ViT) for diagnosing knee osteoarthritis severity. However, due to limited data, this model proved ineffective. To address this, [Wang *et al.*, 2023] replaced ViT's positional embedding with Selective Shuffled Position Embedding (SSPE) and

[*]Corresponding author

[1]Our application is available at: https://osteoga.gamspro.vn/. A demonstration video can be found at: **Google Drive**.

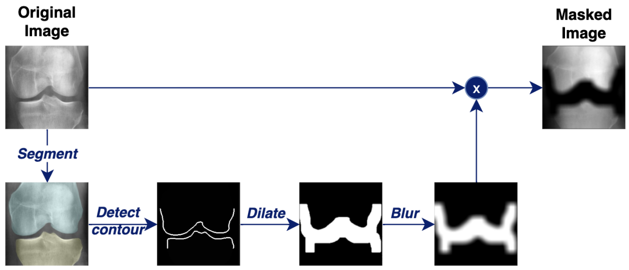

Figure 1: Masking the ROI of a knee joint image using image processing techniques.

employed ROI-Exchange as a data augmentation strategy, improving the model's learning capability while retaining essential knee joint features.

2.2 Anomaly Detection

Anomaly detection involves identifying abnormal patterns in a dataset. In [Siddalingappa and Kanagaraj, 2021], an Autoencoder [Hinton and Salakhutdinov, 2006] was trained to reconstruct normal data accurately. The model's reconstruction error for unseen data was then used to detect anomalies. Another approach involved using a Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014], which trains separate Generator and Discriminator networks to generate high-quality images from random noise. AnoGAN [Schlegl *et al.*, 2017] used gradient descent to determine the optimal latent vector for each input image and measured the difference between the original and generated images to assess anomaly level. Overall, these methods, along with our proposed method, belong to the category of pixel-based anomaly detection, contrasting with the instance-based counterpart.

3 The XGA-Osteo Application

In this section, we introduce the XGA-Osteo application. As an XAI application, XGA-Osteo offers these following major features.

3.1 ROI-Masked Image Reconstruction

The first major feature of XGA-Osteo is the ability to reconstruct knee joint images with a masked region of interest (ROI), which represents the central area of a knee joint. This region is known to contain crucial information about changes and pathologies associated with knee osteoarthritis. To accomplish this, we trained a model called Osteo-GAN.

To train Osteo-GAN, we constructed a dataset called ROI-Masked dataset, as described in Figure 1. The ROI masking process will be automatically performed on knee X-ray images using image processing and computer vision techniques, including segmentation, contour detection, dilation, and blurring. Once the ROI-Masked dataset was obtained, Osteo-GAN was then trained following the process illustrated in Figure 2. Using the ROI-masked healthy images, Osteo-GAN employed adversarial learning to restore original images. Similar to traditional GAN models, adversarial loss was calculated based on the discriminator's ability to distinguish

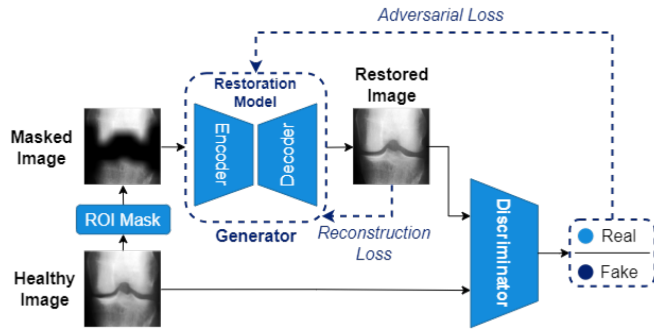

Figure 2: The training process for Osteo-GAN using ROI-masked healthy knee images.

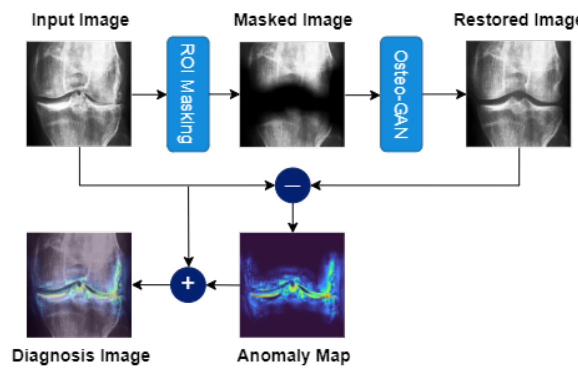

Figure 3: The process of generating an anomaly map from a knee X-ray image.

between real and restored images. Additionally, We employ Mean Absolute Error (MAE) as the reconstruction loss to assess the model's ability to reconstruct the original image from a ROI-masked input. As a result, Osteo-GAN possesses two important characteristics in its reconstruction capability: (i) it only restores the ROI region of the knee X-ray image, and (ii) since Osteo-GAN is trained solely on healthy knee images, it can restore a healthy knee image from a ROI-masked diseased knee image.

3.2 Anomaly Map Generation

Figure 3 presents the process of generating an anomaly map for a knee X-ray image. Essentially, this map is a heatmap that compares pixel-wise differences between the original input image and the restored image from Osteo-GAN. As mentioned earlier, since Osteo-GAN generates a corresponding healthy image from a diseased knee image, the heatmap accurately highlights anomalous points within the ROI region of that image. Consequently, we obtain a diagnostic image that highlights abnormal points, providing an explanation for the model's diagnosis.

Figure 4 illustrates diagnostic images generated from different input cases. As can be observed, for the healthy case, the heat map hardly shows any abnormal points. For the mild case, the heatmap highlights some notable lines around the ROI region, indicating areas of increased density consistent with subchondral sclerosis, a common symptom of the dis-

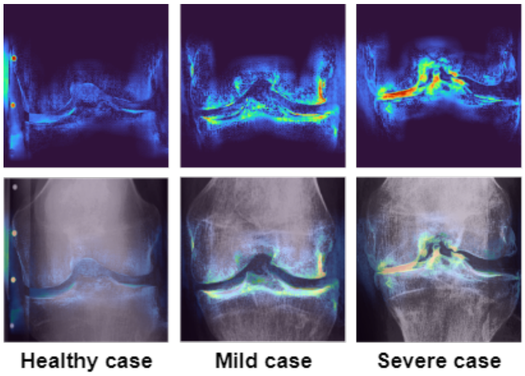

Figure 4: The diagnosis results indicate a high risk of injury to the knee joint.

Figure 5: Comparison of the ability to identify knee joint damage regions between the XGA-Osteo and Grad-CAM method.

ease. Particularly, for the severe case, the heatmap indicates a significant red area that requires attention.

Figure 5 provides a specific analysis of the severe case from Figure 4, from a doctor's perspective. In this case, the red-highlighted region in the heatmap corresponds to joint space narrowing, specifically the highlighted area on the left, as well as the presence of bone spurs, indicated by the smaller area in the center. Consequently, the doctor can provide an explanation for the model's prediction. In other words, the XAI-Osteo application has the capability to provide explainable information that is useful for users. For reference, we also provide the heatmap results provided by Grad-CAM [Selvaraju *et al.*, 2017]. As observed, Grad-CAM can only provide a generic heatmap of the entire region, lacking the ability to highlight specific anomalies.

3.3 Knee Osteoarthritis Severity Diagnosis

The architecture of our classification model is illustrated in Figure 6. In this model, both the original X-ray image and the restored X-ray image are passed through the same backbone model for feature map extraction. Subsequently, we concatenate the Original Feature Map and the Difference Feature Map to generate the Final Feature Map. This process aims to enhance the accuracy of classifying the severity level of the disease in knee X-ray images. The Final Feature Map is then passed through the Global Average Pooling (GAP) layer to reduce computational costs and mitigate overfitting before producing the final diagnostic result.

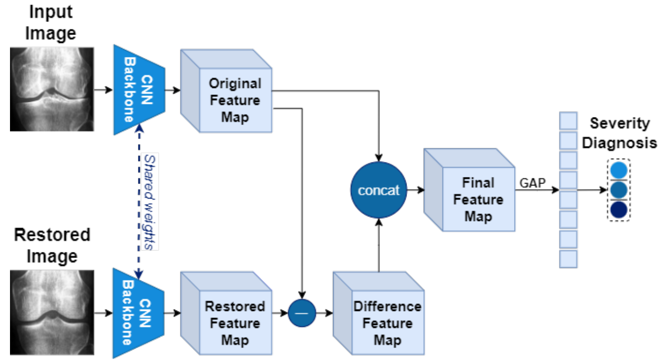

Figure 6: The architecture of the classification model for diagnosing the severity of knee osteoarthritis.

Model	**Accuracy (%)**		**Recall (%)**	
	Base	**Our**	**Base**	**Our**
[Huang *et al.*, 2017]	78.20	76.03	76.79	**78.33**
[He *et al.*, 2016]	78.98	76.51	76.00	**77.00**
[Szegedy *et al.*, 2017]	78.14	75.91	74.63	**75.91**
[Tan and Le, 2021]	77.65	76.69	68.73	**79.00**
[Radosavovic *et al.*, 2020]	78.14	78.02	77.48	**80.67**
[Liu *et al.*, 2022]	77.83	**79.47**	74.43	**81.33**

Table 1: Compare the baseline models with our method using various CNN backbones

4 Experiment

We utilized X-ray image data from the Osteoarthritis Initiative (OAI) project [National Institute of Mental Health, 2001]. Table 1 presents the benchmark results of our method compared to other approaches. The result demonstrates that our method competes with baseline models in terms of accuracy but outperforms them in recall (i.e., better disease detection capability). Moreover, while other prediction methods operate as black boxes, our method is explainable, as discussed earlier. This means it has the ability to provide information to explain its prediction results.

5 Conclusion

In this study, we introduced the XGA-Osteo, an application designed to assist in diagnosing knee osteoarthritis by providing diagnoses and anomaly maps to highlight abnormal regions in knee X-ray images, offering crucial information about the location and severity of the damage. Anomaly maps are extracted using unsupervised anomaly detection, filling the gap in labeled data availability. In the future, we plan to improve the accuracy of our model further to make it more reliable and effective. We expect that this application will be widely used and beneficial for doctors and patients with knee osteoarthritis, improving the diagnosis process.

Acknowledgments

The research is funded by the Vietnam National University Research Project of Type C under grant number C2024-20-32.

References

[Alshareef *et al.*, 2022] Esam Alshareef, Fawzi Omar, Yosra Lamami, Mohamed Milad, Mohamed Eswani, Sedigh Bashir, Salah Bshina, Anas Jakdoum, Asharaf Abourqeeqah, Mohamed Elbasir, and Ellafi.A. Elbahrit. Knee osteoarthritis severity grading using vision transformer. *Journal of Intelligent & Fuzzy Systems*, 43:1–11, 08 2022.

[Georgescu, 2023] Mariana-Iuliana Georgescu. Masked Autoencoders for Unsupervised Anomaly Detection in Medical Images. *Procedia Computer Science*, 225:969–978, 2023. 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023).

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, 2014.

[Gu *et al.*, 2022] Hanxue Gu, Keyu Li, Roy J. Colglazier, Jichen Yang, Michael Lebhar, Jonathan O'Donnell, William A. Jiranek, Richard C. Mather, Rob J. French, Nicholas Said, Jikai Zhang, Christine Park, and Maciej A. Mazurowski. Knee arthritis severity measurement using deep learning: a publicly available algorithm with a multi-institutional validation showing radiologist-level performance, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hinton and Salakhutdinov, 2006] Geoffrey E. Hinton and Ruslan Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[Iqbal *et al.*, 2024] Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. "Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model". In Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang, editors, *Machine Learning in Medical Imaging*, pages 372–381, Cham, 2024. Springer Nature Switzerland.

[Jain *et al.*, 2023] Rohit Kumar Jain, Prasen Kumar Sharma, Sibaji Gaj, Arijit Sur, and Palash Ghosh. Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network. *Multimedia Tools and Applications*, pages 1–18, 2023.

[Lespasio *et al.*, 2017] Michelle J Lespasio, Nicolas S Piuzzi, M Elaine Husni, George F Muschler, AJ Guarino, and Michael A Mont. Knee osteoarthritis: a primer. *The Permanente Journal*, 21, 2017.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.

[National Institute of Mental Health, 2001] National Institute of Mental Health. The Osteoarthritis Initiative. https://nda.nih.gov/oai, 2001. Accessed on: February 2024.

[Radosavovic *et al.*, 2020] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10425–10433, 2020.

[Schlegl *et al.*, 2017] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery". In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen, editors, *Information Processing in Medical Imaging*, pages 146–157, Cham, 2017. Springer International Publishing.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Siddalingappa and Kanagaraj, 2021] Rashmi Siddalingappa and Sekar Kanagaraj. Anomaly Detection on Medical Images using Autoencoder and Convolutional Neural Network. *International Journal of Advanced Computer Science and Applications*, 12(7), 2021.

[Swiecicki *et al.*, 2021] Aleksander Swiecicki, Ning Li, John O'Donnell, Nawar Said, Jun Yang, Richard C Mather, William A Jiranek, and Maciej A Mazurowski. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. *Computers in Biology and Medicine*, 133:104334, Jun 2021.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

[Tan and Le, 2021] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 2021.

[Tiulpin *et al.*, 2018] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.

[Wahyuningrum *et al.*, 2016] Rima Tri Wahyuningrum, Lilik Anifah, I Ketut Eddy Purnama, and Mauridhi Hery Purnomo. A novel hybrid of S2DPCA and SVM for knee osteoarthritis classification. In *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pages 1–5. IEEE, 2016.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition, 2020.

[Wang *et al.*, 2023] Zhe Wang, Aladine Chetouani, and Rachid Jennane. Transformer with Selective Shuffled Position Embedding using ROI-Exchange Strategy for Early Detection of Knee Osteoarthritis. *arXiv preprint arXiv:2304.08364*, 2023.