Diagnostic automatique de la sévérité de la gonarthrose à l'aide de transformeur Swin

Aymen SEKHRI[1, 2] Yassine NASSER[1,4] Marouane TLIBA[1] Mohamed Amine KERKOURI[1] Aladine CHETOUANI[1]
Alessandro BRUNO[3] Ahmed ROUMANE[2] Rachid JENNANE[4]

[1]Laboratoire PRISME, Universite d'Orléans. Orléans, France

[2]Ecole Nationale Supérieure des Télécommunications et des TIC. Oran, Algérie

[3]IULM AI Lab, IULM Libera Università di Lingue e Comunicazione. Milan, Italy

[4]IDP - UMR CNRS 7013, Universite d'Orléans. Orléans, France

Résumé – L'arthrose du genou (gonarthrose) est la maladie articulaire la plus répandue dans le monde. Elle est l'une des principales causes de limitation d'activité et d'incapacité physique chez les personnes âgées. Son diagnostic précoce est essentiel pour une meilleure prise en charge de la maladie et de son suivi. Dans ce papier, nous proposons une approche de classification automatisée pour prédire la sévérité de la gonarthrose à partir d'images radiographiques. Plus précisément, l'architecture proposée est constituée d'un transformeur Swin et d'un en-tête de classification multi-prédictive afin d'améliorer la précision du modèle. L'architecture proposée a été entraînée et évaluée sur deux grandes bases de données publiques. Les résultats obtenus montrent le potentiel et la faisabilité de l'approche proposée.

Abstract – Knee OsteoArthritis (OA), is a common disease that affects the articular cartilage of the knee joint, causing pain and stiffness. Over time, the disease can worsen and lead to more severe symptoms and complications, such as major disability, with loss of mobility. Early detection and diagnosis of knee OA is essential for effective clinical intervention and management. In this paper, we propose an automated classification approach to predict the severity of knee OA using the Swin Transformer. We introduce a multi-predictive head architecture using multi-layer perceptron classifiers to improve the accuracy of our model. Our model was evaluated on two publicly available radiographic database. Obtained results show the potential and the feasibility of the proposed approach.

1 Introduction

L'arthrose du genou ou gonarthrose est une maladie articulaire répandue qui affecte principalement le cartilage articulaire du genou, provoquant des douleurs, des raideurs et une mobilité réduite. Il s'agit d'une affection qui peut entraîner des changements osseux et l'usure du cartilage, ce qui conduit finalement à une chirurgie de remplacement total de l'articulation. La gonarthrose est très répandue chez les personnes âgées, les personnes en surpoids et celles qui ont un mode de vie sédentaire. En raison de sa nature progressive, une détection et un diagnostic précoce sont essentiels pour une intervention clinique et une prise en charge efficaces. [11, 13]

La gonarthrose peut être diagnostiquée sur la base de résultats cliniques et radiographiques. Le rétrécissement de l'espace articulaire, la formation d'ostéophytes et la sclérose sont les principales caractéristiques pathologiques de la gonarthrose [11, 8] qui peuvent être diagnostiquées à l'aide de radiographies [12]. Bien que diverses techniques telles que l'imagerie par résonance magnétique, la tomodensitométrie et l'échographie aient été introduites pour le diagnostic de la gonarthrose, la radiographie reste la méthode la plus utilisée pour le diagnostic initial en raison de son accessibilité, son faible coût et de son utilisation généralisée. Les bases de données publiques, telles que : OsteoArthritis Initiative (OAI, `https://nda.nih.gov/oai/`) et Multicenter Osteoarthritis Study (MOST, `https://most.ucsf.edu/`), fournissent des radiographies avec les grades de Kellgren et Lawrence (KL) [5], qui constituent des ressources précieuses pour la recherche sur l'arthrose du genou.

Dans ce travail, nous examinons tout d'abord l'utilisation du transformeur Swin pour la prédiction de la sévérité de la gonarthrose à partir d'images radiographiques. La méthode proposée utilise le réseau transformeur Swin [6] comme réseau principal dans le but d'extraire des caractéristiques de haut niveau pour détecter les changements induits par la gonarthrose. Ensuite, nous introduisons un en-tête de classification multi-prédictif afin de faire face au problème de la grande similarité entre les différents grades de la gonarthrose [10, 9] (voir figure 1). De plus, pour réduire le problème de décalage entre les données des deux bases de données OAI et MOST, nous avons testé plusieurs méthodes d'apprentissage pour trouver celle qui donne les meilleurs résultats équilibrés.

Le reste de ce document est organisé comme suit : la méthode proposée est décrite dans la section 2. Les résultats expérimentaux obtenus sont présentés dans la section 3. La conclusion et les perspectives sont données dans la section 4.

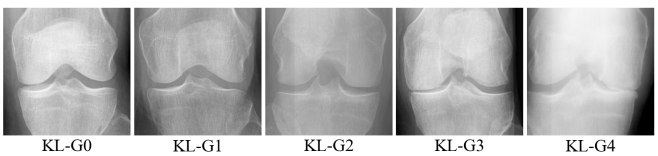

FIGURE 1 : Quelques échantillons de radiographies de la gonarthrose. KL-G0 : genou sain sans arthrose, KL-G1 : arthrose douteuse, KL-G2 : arthrose minime, KL-G3 : arthrose modérée et KL-G4 : arthrose sévère.

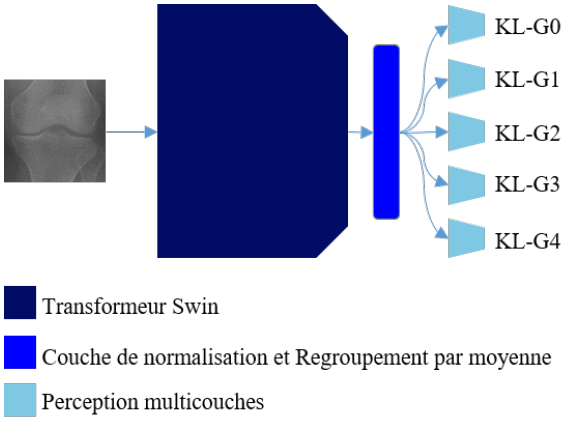

FIGURE 2 : Architecture du modèle : Le transformeur swin avec une architecture de tête de prédiction multiple

2 Methode proposée

La méthode proposée utilise le modèle de transformeur Swin comme réseau dorsal pour notre architecture. Nous avons attaché des réseaux de perceptrons multicouches (multi-layer perceptron, MLP) pour chaque classe au transformeur Swin afin d'augmenter la capacité du modèle à mieux séparer les images similaires appartenant à des classes différentes. L'architecture proposée est illustrée par la figure 2.

2.1 Les transformeurs de vision

Le transformeur de vision est une architecture de réseau de neurones qui s'appuie sur l'architecture originale du modèle transformeur pour la traduction automatique [15], la principale différence est que la partie décodeur du transformeur est supprimée.

Afin de traiter une image $X \in R^{H\times W\times 3}$, on crée N patchs $x_i \in R^{p\times p\times 3}$, qu'on transforme en vecteurs en utilisant une couche linéaire et des encodeurs de position (i.e. *embeddings*) pour préserver l'information spatiale. Les vecteurs $1D$ sont ensuite envoyés en entrée d'un encodeur standard qui utilise des blocs d'auto-attention pour générer une séquence de vecteur de caractéristiques. Cette séquence est ensuite classifiée par des MLP finaux [4].

2.2 Le transformeur Swin

Le transformeur Swin est l'un des modèles de transformeurs les plus efficaces qui ait été proposé pour surmonter les difficultés liées au transfert des performances des transformeurs du domaine linguistique à celui du visuel [6]. Le transformeur Swin utilise une structure hiérarchique efficace pour traiter des entrées de différentes tailles. Il exploite également des fenêtres décalées et des couches de fusion pour traiter des objets à différentes échelles, ce qui lui permet d'obtenir les meilleurs résultats sur divers benchmarks de vision par ordinateur.

Le modèle de base comporte quatre étapes. Premièrement, le réseau prend l'image d'entrée de taille $H \times W \times 3$ et la partitionne en patchs adjacents de taille 4×4, soit au total $\frac{H}{4} \times \frac{W}{4}$ jetons (i.e. *tokens*) de taille $4 \times 4 \times 3 = 48$. Ces jetons sont ensuite transmis à une couche d'encodage de position qui projette ces jetons dans une dimension C. Cette couche d'encodage de position avec deux blocs successifs de transformeur Swin constitue la première étape du modèle. Le premier bloc du transformeur Swin utilise un mécanisme basé sur des fenêtres d'auto-attention multi-têtes (W-MSA : *Window-based Multi-head Self-Attention*), dans lequel l'auto-attention est calculée uniquement entre les patchs situés à l'intérieur de la même fenêtre (où chaque fenêtre contient $M \times M$ patchs), tandis que le second bloc utilise le concept de l'auto-attention multi-têtes à fenêtres décalées (SW-MSA : : *Shifted-Window Multi-head Self-Attention*) dans laquelle les fenêtres de partitionnement sont déplacées de $|\frac{M}{2}| \times |\frac{M}{2}|$ pixels par rapport aux fenêtres de partitionnement standard utiliser dans le bloc précédent. Cette nouvelle approche proposée avec les transformeurs Swin a pour objectif de créer davantage de relations entre les patchs voisins qui n'étaient pas situés sur la même fenêtre auparavant. Les deux blocs consécutifs dus transformeur Swin sont calculés comme suit :

$$\begin{aligned} \hat{z}^l &= W - MSA(LN(z^{l-1})) + z^{l-1} \\ z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l \\ \hat{z}^{l+1} &= SW - MSA(LN(z^l)) + z^l \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{1}$$

où $\hat{z}^l$ et z^l représentent respectivement les résultats du module (S)W-MSA et du module MLP pour un bloc l.

Dans la deuxième étape, une couche de fusion de patchs est appliquée pour regrouper chaque 2×2 patch voisin en un seul patch de longueur $4C$, réduisant ainsi le nombre de patchs à $\frac{H}{8} \times \frac{W}{8}$. Ces patchs sont ensuite projetés linéairement à une dimension de taille $2C$ et transmis à deux blocs de transformeur Swin comme dans la première étape.

Ce processus est répété dans la troisième étape, en utilisant 18 blocs de transformeur Swin pour produire $\frac{H}{16} \times \frac{W}{16}$ patchs de longueur $4C$. Finalement, à la quatrième étape, deux blocs de transformeur Swin sont utilisés pour produire des patchs $\frac{H}{32} \times \frac{W}{32}$ d'une longueur de $8C$. Ces étapes consécutives ont produit conjointement une représentation hiérarchique comme celles des réseaux convolutionnels classiques.

2.3 Architecture de tête à prédiction multiple

Pour tenir compte de la grande similarité entre les radiographies de sévérité proche (Figure 1), nous avons utilisé un en-tête de classification multi-prédictif, cela signifie que nous considérons la prédiction de chaque grade comme un problème de classification binaire, de sorte que chaque tête MLP prédit si le grade existe ou non. Notre modèle final consiste en un encodeur Swin-B de base, avec $C = 128$ et $2, 2, 18, 2$ blocs de transformeur Swin, suivi par des couches de normalisation et regroupement par moyenne (i.e. *average pooling*), afin de produire un vecteur de caractéristiques final d'une taille de 1024. Ce vecteur est transmis par la suite à 5 blocs de classifications, un pour chaque classe considérée. Chaque bloc contient 3 couches MLP de taille $384, 48, 48$, et une dernière couche avec un seul neurone pour prédire le grade final.

La classe prédite correspond au score de confiance maximale issue des multiples têtes après l'application de la non linéarité de la fonction Sigmoïde.

2.4 Implémentation

Afin d'entraîner le modèle, nous avons utilisé l'optimiseur AdamW [7] avec un taux d'apprentissage de 3e-5, une décroissance des poids de 0,05, un epsilon de 1e-8 et des bêtas de (0,9, 0,999) pour ajuster les poids. Nous avons entraîné le modèle sur des batches de 32 images pendant une période de 300 epochs. Nous avons implémenté le code en PyTorch et utilisé une carte GPU NVIDIA RTX A4000 avec 16 Go de mémoire pour accélérer le processus d'entraînement.

3 Résultats expérimentaux

3.1 Bases de données

Les données utilisées dans notre étude proviennent de deux bases de données accessibles au public : MOST et OAI. La base MOST est constituée de 18 376 images de genoux segmentés de la même manière que [14]. Dans cette étude, nous avons partitionné les données en 3 ensembles pour l'entraînement, la validation et la phase de test avec un ratio 6 :1 :3. Pour la base de données OAI contenant 8260 images de genoux prétraités [3], nous l'avons partitionné de manière aléatoire avec un ratio de 7 :1 :2, respectivement pour l'entraînement, la validation et le test. Nous avons considéré les 5 grades KL allant de KL-G0 (genou normal) à KL-G4 (gonarthrose sévère). Pour résoudre le problème des données limitées et de l'overfitting, nous avons appliqué diverses techniques d'augmentation des données telles que la rotation de 15 degrés, la translation, la mise à l'échelle, le retournement horizontal aléatoire et l'ajustement du contraste avec un facteur de 0,3. Ces techniques ont déjà été utilisées dans des travaux similaires pour améliorer les performances des modèles d'apprentissage profond sur des tâches de classification d'images.

3.2 Résultats de la classification

Afin de tester l'efficacité du modèle proposé pour prédire les différentes sévérités de la gonarthrose et pour trouver la meilleure façon d'entrainer le modèle et ainsi améliorer sa capacité de généralisation, nous avons mené quatre expériences.

Dans la première expérience, nous avons entraîné notre modèle exclusivement sur la base MOST et évalué sa capacité de généralisation sur la base OAI.

Dans la deuxième expérience, nous avons entraîné notre modèle sur l'ensemble d'entraînement à la fois sur les deux bases OAI et MOST simultanément. Cette expérience permet d'améliorer les performances du modèle en terme de généralisation en exploitant les informations des deux bases.

Dans la troisième expérience, nous avons tenté de réduire la différence entre les bases OAI et MOST. Pour ce faire, nous avons entraîné le modèle entier sur l'ensemble d'entraînement de la base MOST, puis nous avons figé les couches de classification et entraîné uniquement le backbone de notre modèle (i.e. le transformeur Swin) sur la base OAI. Cette expérience vise à déterminer si le transfert des connaissances du classifieur formé sur la base MOST peut améliorer les performances tout en comblant le décalage entre les deux bases.

Enfin, la quatrième expérience a consisté à utiliser une seule tête ou bloque de perceptron multicouche pour prédire tous les grades simultanément. Cette expérience a pour objectif d'évaluer l'impact de l'utilisation d'une architecture de tête de prédiction multiple.

Exp.	MOST		OAI	
	Acc (%) ↑	F1 ↑	Acc (%) ↑	F1 ↑
1	75.43	0.714	62.86	0.615
2	73.13	0.684	66.85	0.657
3	73.25	0.667	**70.17**	**0.671**
4	71.93	0.622	67.15	0.615

TABLE 1 : Comparaison des quatre expériences en termes de performance de précision et de score F1 sur les bases OAI et MOST.

Les résultats obtenus sont présentés dans le tableau 1. L'expérience 1 montre que le modèle entraîné uniquement sur la base MOST a une capacité de généralisation limitée, car la précision de classification est plus faible sur la base OAI. Cela est essentiellement dû au décalage entre les deux distributions des deux bases de données.

Lors de la deuxième expérience avec les deux bases de données, une amélioration de la précision sur la base OAI est obtenue avec une réduction de la précision sur la base MOST. Ainsi, cette stratégie a réduit le décalage entre les deux bases.

L'objectif de l'expérience 3 est de réduire l'écart entre les deux bases tout en conservant les capacités du classifieur formé sur la base MOST. En entraînant uniquement le backbone sur l'ensemble de données OAI après avoir entraîné l'ensemble du modèle sur MOST, nous observons une augmentation conséquente de la précision sur la base OAI avec un taux de 70.17%.

La quatrième expérience vise à montrer la pertinence d'utiliser un en-tête de classification multi-prédictif. Les résultats obtenus montrent une diminution de la précision, particulièrement une difficulté importante à prédire le score KL-G1 au test OAI.

3.3 Comparaison avec les méthodes de l'état de l'art

Le tableau 2 présente une comparaison des résultats obtenus avec l'état de l'art. Parmi les méthodes considérées, il y a celles qui utilisent des fonctions de perte différentes telles que la perte ordinale au lieu de la perte d'entropie croisée qui introduit des poids de pénalité entre les grades prédits et les grades réels [3]. D'autres travaux [1] ont utilisé une combinaison de pertes de classification et de régression en tenant compte de la nature progressive de la maladie. Un travail précédent [14] a utilisé un réseau Siamois en considérant deux régions d'intérêt pour tirer parti de la symétrie de l'image. Cette approche a permis à l'architecture d'apprendre des poids identiques pour les deux côtés de l'image ce qui s'est traduit par de meilleures performances. Une autre approche prometteuse [16] a utilisé un concept similaire aux transformeurs de vision en employant une architecture ResNet50 pour extraire des cartes de caractéristiques spatiales de la même région locale de l'image. Ensuite, ces cartes ont été vectorisées avec des encodeurs de position et sont passés par 12 blocs de transformeurs visuels pour la classification.

Il est important de reconnaître que les méthodes utilisées dans ces études ont été apprises différemment. Plus précisément, certaines méthodes ont utilisé exclusivement la base OAI comme ensemble d'apprentissage, d'autres ont utilisé exclusivement l'ensemble de la base MOST, tandis que d'autres ont utilisé les deux bases. Cette diversité dans l'apprentissage peut avoir un impact sur la performance globale, et doit donc être considérée avec attention lors de l'interprétation des résultats.

L'approche proposée a obtenu les meilleures performances parmi les méthodes comparées. Ainsi, l'utilisation de l'architecture de transformeur Swin permet d'améliorer les performances comparées à d'autres méthodes en apprenant des caractéristiques sémantiques de manière hiérarchique, ce qui est particulièrement utile pour capturer des informations pertinentes sur la progression et la sévérité de la gonarthrose.

4 Conclusion

Dans ce papier, nous avons proposé une nouvelle méthode pour prédire la sévérité de la gonarthrose à partir d'images radiographiques en utilisant le réseau transformeur Swin. Nos résultats démontrent que cette méthode atteint des performances de pointe sur l'ensemble de données de référence OAI, surpassant

Method	Acc (%) ↑	F1 ↑
Antony et al. 2016 [2]	53.40	0.43
Antony et al. 2017 [1]	63.60	0.59
Ordinal Loss (Vgg19) [3]	69.60	-
Ordinal Loss (ResNet50) [3]	66.20	-
Ordinal Loss (ResNet101) [3]	65.50	-
Siamese net [14]	66.71	-
Yifan et al. [16]	69.18	-
Le Notre (config. 3)	**70.17**	**0.67**

TABLE 2 : Résultats obtenus avec la base de données OAI

de manière significative les méthodes existantes. Nous avons montré que le réseau de transformeurs Swin est efficace pour extraire des informations pertinentes sur la gonarthrose, qui peuvent être utilisées pour détecter les changements induits par la gonarthrose dans le tissu osseux. En outre, notre tête de classification multi-prédictive améliore considérablement la précision du modèle et aide à réduire la similarité entre les caractéristiques des grades proches. Les perspectives à ce travail peuvent concerner d'autres modalités d'imagerie telles que l'IRM, tout en explorant les données cliniques et démographiques, pour améliorer encore plus la prédiction de la sévérité de la gonarthrose.

Références

[1] Joseph ANTONY, Kevin MCGUINNESS, Kieran MORAN et Noel E O'CONNOR : Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. *In Machine Learning and Data Mining in Pattern Recognition : 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, pages 376–390. Springer, 2017.

[2] Joseph ANTONY, Kevin MCGUINNESS, Noel E O'CONNOR et Kieran MORAN : Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. *In 2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE, 2016.

[3] Pingjun CHEN, Linlin GAO, Xiaoshuang SHI, Kyle ALLEN et Lin YANG : Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019.

[4] Alexey DOSOVITSKIY, Lucas BEYER, Alexander KOLESNIKOV, Dirk WEISSENBORN, Xiaohua ZHAI, Thomas UNTERTHINER, Mostafa DEHGHANI, Matthias MINDERER, Georg HEIGOLD, Sylvain GELLY *et al.* : An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[5] Jonas H KELLGREN et JS1006995 LAWRENCE : Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957.

[6] Ze LIU, Yutong LIN, Yue CAO, Han HU, Yixuan WEI, Zheng ZHANG, Stephen LIN et Baining GUO : Swin transformer : Hierarchical vision transformer using shifted windows. *In Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[7] Ilya LOSHCHILOV et Frank HUTTER : Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*, 2017.

[8] Anne CA MARIJNISSEN, Koen L VINCKEN, Petra AJM VOS, DBF SARIS, MA VIERGEVER, JWJ BIJLSMA, LW BARTELS et FPJG LAFEBER : Knee images digital analysis (kida) : a novel method to quantify individual radiographic features of knee osteoarthritis in detail. *Osteoarthritis and cartilage*, 16(2):234–243, 2008.

[9] Yassine NASSER, Mohammed El HASSOUNI et Rachid JENNANE : Discriminative deep neural network for predicting knee osteoarthritis in early stage. *In Predictive Intelligence in Medicine : 5th International Workshop, PRIME 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 126–136. Springer, 2022.

[10] Yassine NASSER, Rachid JENNANE, Aladine CHETOUANI, Eric LESPESSAILLES et Mohammed EL HASSOUNI : Discriminative regularized auto-encoder for early detection of knee osteoarthritis : data from the osteoarthritis initiative. *IEEE transactions on medical imaging*, 39(9):2976–2984, 2020.

[11] H OKA, S MURAKI, T AKUNE, A MABUCHI, T SUZUKI, H YOSHIDA, S YAMAMOTO, K NAKAMURA, N YOSHIMURA et H KAWAGUCHI : Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage*, 16(11):1300–1306, 2008.

[12] Hee-Jin PARK, Sam Soo KIM, So-Yeon LEE, Noh-Hyuck PARK, Ji-Yeon PARK, Yoon-Jung CHOI et Hyun-Jun JEON : A practical mri grading system for osteoarthritis of the knee : association with kellgren–lawrence radiographic scores. *European journal of radiology*, 82(1):112–117, 2013.

[13] Lior SHAMIR, Shari M LING, William SCOTT, Marc HOCHBERG, Luigi FERRUCCI et Ilya G GOLDBERG : Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage*, 17(10):1307–1312, 2009.

[14] Aleksei TIULPIN, Jérôme THEVENOT, Esa RAHTU, Petri LEHENKARI et Simo SAARAKKALA : Automatic knee osteoarthritis diagnosis from plain radiographs : a deep learning-based approach. *Scientific reports*, 8(1):1–10, 2018.

[15] Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N GOMEZ, Łukasz KAISER et Illia POLOSUKHIN : Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Yifan WANG, Xianan WANG, Tianning GAO, Le DU et Wei LIU : An automatic knee osteoarthritis diagnosis method based on deep learning : data from the osteoarthritis initiative. *Journal of Healthcare Engineering*, 2021:1–10, 2021.