Enhanced deep neural networks for early diagnosis of knee osteoarthritis

Yassine Nasser

ÉCOLE DOCTORALE MIPTIS

INSTITUT DENIS POISSON / LABORATOIRE DE RECHERCHE EN INFORMATIQUE ET TÉLÉCOMMUNICATION

THÈSE EN COTUTELLE INTERNATIONALE présentée par :

Yassine NASSER

soutenue le : 20 mars 2023

pour obtenir le grade de : Docteur de l'Université d'Orléans et de l'Université Mohammed V de Rabat

Discipline/Spécialité : Sciences et Technologies Industrielles - Informatique

ENHANCED DEEP NEURAL NETWORKS FOR EARLY DIAGNOSIS OF KNEE OSTEOARTHRITIS

THÈSE dirigée par :

M. JENNANE Rachid	PU, IDP, Université d'Orléans
M. EL HASSOUNI Mohammed	PES, FLSH, Univ. Mohammed V de Rabat

RAPPORTEURS :

M. ADIB Abdellah	PES, FSTM, Univ. Hassan II de Casablanca
Mme. CHAPPARD Christine	PH, INSERM, Université Paris Diderot
M. El HAZITI Mohamed	PES, EST, Univ. Mohammed V de Rabat

JURY :

M. ABRAHAM Romain	PU, IDP, Univ. d'Orléans, Président du jury
M. ADIB Abdellah	PES, FSTM, Univ. Hassan II de Casablanca
Mme. CHAPPARD Christine	PH, INSERM, Université Paris Diderot
M. El HAZITI Mohamed	PES, EST, Univ. Mohammed V de Rabat
M. HANS Didier	PA, Université de Lausanne
M. EL HASSOUNI Mohammed	PES, FLSH, Univ. Mohammed V de Rabat
M. JENNANE Rachid	PU, IDP, Université d'Orléans

Résumé

L'arthrose du genou, ou la gonarthrose, est l'une des principales causes de handicaps physiques dans le monde associés à un fardeau personnel et socio-économiques importantes. Il existe un besoin considérable pour le développement des méthodes automatisées pour aider au diagnostic précoce de la gonarthrose. Au cours des dernières années, les modèles d'apprentissage profonds ont suscité un intérêt remarquable de la part de la communauté scientifique et ont remporté un grand succès dans diverses applications d'imagerie médicale. Cette thèse vise à développer des modèles basés sur l'apprentissage profond pour le diagnostic automatique de la gonarthrose à l'aide d'images radiographiques. Plusieurs méthodes d'évaluation de la gravité de la gonarthrose sont testées et des nouvelles sont introduites.

Tout d'abord, nous avons investigué l'étape d'apprentissage des caractéristiques en tant que composante importante du système de classification afin d'extraire les caractéristiques discriminantes les plus utiles à partir d'images radiographiques. Pour ce faire, nous introduisons une nouvelle architecture, appelée Discriminative Regularized Auto-Encoder(DRAE), basée sur les réseaux de neurones type Auto-Encodeurs. Le but de ce modèle est de séparer les différentes classes d'images (arthrosiques et saines) en minimisant la distance entre les sujets de même classe (intra-classes) et en maximisant la distance entre les sujets de classes différentes (inter-classes).

Ensuite, afin de mieux détecter les signes précoces de la gonarthrose, nous proposons d'intégrer la régularisation discriminante proposée dans le processus d'apprentissage de réseaux de neurones convolutifs (CNN). En agissant ainsi, nous améliorons le potentiel du CNN standard à traiter des données présentant de fortes similitudes inter-classes ou de fortes variations intra-classes. Tandis que le DRAE proposé se concentre sur les informations de texture dans la radiographie, ce deuxième modèle, appelé Discriminative Convolutional Neural Network (DCNN), utilise la zone distale globale du genou et analyse à la fois les représentations de texture et de forme. Pour aller plus loin dans l'apprentissage des caractéristiques discriminantes et l'exploitation d'informations types forme et texture, nous avons proposé : (i) d'améliorer l'analyse de la texture en ajoutant un nouveau bloc à l'architecture DCNN et (ii) améliorer la perte discriminative pour l'adapter aux tâches de classification multi-classes. Le modèle résultant, appelé Discriminative Shape-Texture Convolutional Neural Network (DST-CNN), a fourni une meilleure performance de classification par rapport aux modèles d'apprentissage existant dans la littérature.

Abstract

Knee Osteoarthritis (OA) is one of the most frequent causes of physical disability worldwide and is associated with significant personal and socioeconomic burdens. There is a considerable need to develop automated methods for early diagnosis of knee OA. Over the last few years, Deep Learning (DL) models have gained remarkable attention from the computer vision research community and achieved great success in various medical imaging applications. This thesis aimed to develop DL-based models for fully automatic knee OA diagnosis using radiographic images. In this thesis, several methods for knee OA severity assessment and OA prediction are evaluated, and new methods are introduced.

First, we focus on the feature-learning step as a crucial component of the classification system to learn and extract the most useful discriminative features from X-ray images. To do so, we introduce a novel autoencoder-based architecture called Discriminative Regularized Auto-encoder (DRAE). The goal is to maximize the distance between class features by minimizing the intraclass distance and maximizing the interclass distance. Then, we propose incorporating the proposed discriminative regularization in the standard Convolutional Neural Network (CNN) learning process to improve the early detection of knee OA. By doing so, we reduce the inability of CNNs to handle data with high inter-class similarities or high intra-class variations. While the proposed DRAE focuses on the texture information in the radiography under the tibial plateau, the second proposed learning model, called Discriminative Convolutional Neural Network (DCNN), uses the overall distal area of the knee and exploits both texture and shape representations. To further learn discriminative features and exploit shape and texture, we propose : (i) enhancing texture analysis by adding a new block to the DCNN architecture and (ii) improving the proposed discriminative loss to fit with multi-class classification tasks. The resulting model, called Discriminative Shape-Texture Convolutional Neural Network (DST-CNN), enables better and well-balanced classification performance than existing State-of-the-Art (SoA) models.

Keywords : Deep learning, Feature learning, Auto-Encoder, CNNs, Discriminative regularization, knee osteoarthritis, plain radiography, X-ray.

Acknowledgments

I would like to begin this thesis by express my deepest gratitude and appreciation to all those who have supported me throughout my research journey and contributed to the successful completion of my thesis. I feel incredibly fortunate to have had this adventure, which proved to be a remarkable mix of scientific exploration and personal development.

First and foremost, I extend my heartfelt appreciation to my supervisors Rachid JENNANE and Mohammed EL HASSOUNI, for their valuable guidance, expertise, and encouragement throughout my research journey. Their constant support, motivation and constructive feedback was instrumental in shaping the direction and progress of my work. I am truly grateful for their mentorship as well as for their great patience.

I want to express my sincerest thanks to each member of my exceptional jury for their invaluable time and expertise in evaluating my work. I want to extend my sincere gratitude to Christine CHAPPARD, Abdellah ADIB, and Mohamed El HAZITI for graciously agreeing to review my work and for the stimulating scientific discussions we had. I am also grateful to Romain ABRAHAM for expertly presiding over my defense and to Didier HANS for her thoughtful evaluation. The members of my jury contributed greatly to making this day an enjoyable and memorable experience.

I am also indebted to the members of the LRIT laboratory and all the professors of the department of physics and computer science at the Mohammed V University in Rabat, and the members of the I3MTO, IDP and PRISME laboratories at the University of Orléans, for providing me with the academic background and knowledge and the research environment and facilities that allowed me to carry out this research. I would especially like to thank Salma MOULINE for her financial assistance and Aladine CHETOUANI for his technical and scientific help throughout my thesis.

I do not forget my colleagues and friends who have supported me in various ways during the challenging times of my research : Safae Azzakhnini, Abdelouahid Bentamou, Soufiane Faieq, Mohamed El Yafrani,

Naima Otberdout, Mohamed Hafri, Evelyn Gutierrez, Mohamed amine Kerkouri, and Marouane Tliba.

The journey to completing a thesis is a challenging and often overwhelming one, which requires the support and encouragement of loved ones. I would like to express my heartfelt appreciation to my family for their unwavering love and support throughout this adventure. Their belief in me has been an essential source of strength, and I am forever grateful. My parents, Latifa Boussof and Mohamed Nasser, have been a constant source of love and support, providing me with the emotional and education necessary to navigate the challenges of life. I want to express also my gratitude to my brother and sister for their unwavering support and encouragement throughout my studies. Last but not least, my wife, Hala Hafidi, has been a constant source of support, encouragement, and love. I owe her a debt of gratitude for her unwavering patience and understanding, which have been indispensable in allowing me to pursue my dreams and goals.

Résumé substantiel

Contexte

L'ARTHROSE du genou, ou la gonarthrose, est l'une des principales causes de handicap physique dans le monde, associée à un fardeau personnel et socio-économique important. Il existe un besoin considérable de développer des méthodes automatisées pour aider au diagnostic précoce de la gonarthrose. Au cours des dernières années, l'intelligence artificielle (IA) a connu une progression fulgurante. Cette avancée se manifeste par son impressionnante réussite dans de multiples applications relevant de divers domaines, tels que la santé, l'économie, l'éducation, les véhicules autonomes et les médias. La vision par ordinateur est un domaine intéressant de l'IA qui vise à doter les ordinateurs ou les machines de capacité visuelle, c'est-à-dire à leur permettre d'analyser et de comprendre automatiquement une image à l'aide d'algorithmes informatiques. Ce domaine recouvre différents problèmes, tels que la détection, la segmentation, la reconnaissance, l'estimation du mouvement et la restauration d'images. Alors que la vision par ordinateur a pour objectif de relever le défi de la compréhension d'une image, l'apprentissage profond (deep learning) s'attache à comprendre la grande quantité de données disponibles. Les modèles d'apprentissage profonds ont suscité un intérêt remarquable de la part de la communauté scientifique et ont remporté un grand succès dans diverses applications d'imagerie médicale. Cette thèse vise à développer des modèles basés sur l'apprentissage profond pour le diagnostic automatique de la gonarthrose à l'aide d'images radiographiques.

L'arthrose (OA) est la forme la plus courante de maladie articulaire. C'est une maladie qui entraîne la dégradation irréversible du cartilage qui recouvre les extrémités des os au niveau des articulations. Elle se manifeste souvent par une raideur, un gonflement, une douleur et un craquement lors des mouvements. L'arthrose apparaît généralement vers 50 ans et touche davantage les femmes que les hommes. Environ 21 millions d'adultes américains ont reçu un diagnostic d'arthrose par un médecin, basé sur une combinaison des signes articulaires et d'anomalies radiologiques. Cependant, beaucoup de patients ont une maladie non détectée ou peu sévère. La prévalence de l'arthrose dans la population est difficile à évaluer, car il existe une grande variabilité entre les changements radiologiques et les symptômes des personnes atteintes. Selon l'organisation mondiale de la santé, $9,6\%$ des hommes et 18% des femmes de plus de 60 ans souffrent d'arthrose symptomatique. On estime que 80% des personnes atteintes d'arthrose auront des difficultés

à se mouvoir et que 25% ne pourront pas accomplir les principales activités quotidiennes de leur vie. On prévoit que les affections dégénératives des articulations, comme l'arthrose, toucheront au moins 130 millions de personnes dans le monde d'ici 2050. Cette maladie représente donc un énorme coût pour les systèmes de santé (représentant 1 à $2,5\%$ du produit national brut dans les pays occidentaux) et ce coût devrait doubler d'ici 2030. De plus, aucun traitement n'est capable de stopper la dégradation des structures articulaires responsables de l'évolution de l'arthrose du genou. La plupart des traitements actuels visent seulement à soulager la douleur et à limiter ou à éviter le handicap dû à la destruction des os et du cartilage.

La gonarthrose (Knee OA) est causée par une dégradation du cartilage articulaire du genou et des modifications de la micro-architecture osseuse. La maladie est reconnue comme la principale cause de perte de mobilité chez les personnes âgées et est maintenant considérée comme un facteur de risque indépendant de mortalité accrue. Les causes exactes de l'arthrose du genou ne sont pas entièrement élucidées, mais il existe plusieurs facteurs qui peuvent favoriser son développement :

- L'âge : avec le vieillissement, le risque de développer une arthrose du genou augmente. Cela s'explique par le fait que le cartilage du genou s'use naturellement avec le temps.
- Le sexe : les femmes sont plus susceptibles de développer une arthrose du genou que les hommes.
- La génétique : il existe des preuves que la génétique joue un rôle dans le développement de l'arthrose du genou. Les personnes ayant des antécédents familiaux de la maladie sont plus susceptibles de la développer elles-mêmes.
- L'obésité : le surpoids exerce une pression supplémentaire sur l'articulation du genou, ce qui peut accélérer la dégénérescence du cartilage.
- Les blessures articulaires : des blessures antérieures à l'articulation du genou, comme une rupture d'un ligament ou d'un ménisque, peuvent augmenter le risque de développer une arthrose du genou.
- Le stress répétitif : les métiers ou les activités qui nécessitent un stress répétitif sur l'articulation du genou, comme le travail dans le bâtiment ou la course à pied, peuvent augmenter le risque de développer une arthrose du genou.

La sévérité de la gonarthrose peut varier considérablement, de légère à sévère. Les symptômes courants comprennent la douleur, la raideur, le gonflement et une diminution de l'amplitude de mouvement du genou. L'arthrose du genou est généralement évaluée par des examens d'imagerie, tels que la radiographie (rayons X), l'imagerie par résonance magnétique (IRM) ou la tomodensitométrie (TDM), qui peuvent montrer la plupart des signes caractéristiques de l'arthrose du genou (voir figure 2.2). La radiographie est considérée comme le standard de référence pour l'évaluation de l'arthrose du genou en raison de sa large accessibilité, de son faible coût et de sa sécurité. Les signes caractéristiques de l'arthrose du genou

comprennent le rétrécissement de l'espace articulaire, la perte de cartilage, les excroissances osseuses (ostéophytes), la sclérose sous-chondrale et d'autres changements dans l'articulation, illustrés sur la figure 2.1. En se basant sur ces caractéristiques, Kellgren et Lawrence (KL) ont divisé la gravité de l'arthrose du genou en cinq stades (voir figure 2.3) :

- Normal (grade 0) : il n'y a pas de rétrécissement de l'espace articulaire ni d'ostéophytes. L'articulation du genou apparaît normale sur la radiographie.
- Douteux (grade 1) : il y a un rétrécissement minime de l'espace articulaire et/ou un petit ostéophyte est présent. Ce stade est souvent asymptomatique et le patient peut ne présenter aucun symptôme.
- Léger (grade 2) : il y a un rétrécissement léger de l'espace articulaire et/ou des petits ostéophytes sont présents. Les patients peuvent ressentir une douleur et une raideur légères, surtout après des périodes prolongées d'activité ou après être restés assis longtemps.
- Modéré (grade 3) : il y a un rétrécissement modéré de l'espace articulaire et/ou des ostéophytes de taille moyenne sont présents. Les patients peuvent ressentir une douleur et une raideur plus fréquentes, surtout lors des activités sollicitant le poids du corps et peuvent avoir des limitations dans leur capacité à effectuer les activités quotidiennes.
- Sévère (grade 4) : il y a un rétrécissement sévère de l'espace articulaire et/ou des ostéophytes de grande taille sont présents. Les patients peuvent ressentir une douleur et une raideur constantes, même au repos et peuvent avoir des limitations importantes dans leur capacité à effectuer les activités quotidiennes. Dans les cas sévères, une déformation articulaire peut également être présente.

  Cependant, l'évolution des caractéristiques impliquées dans l'arthrose du genou est continue ; par conséquent, la classification en grade distinct est souvent laissée à l'appréciation subjective de l'annotateur11. Cela introduit de la subjectivité/de l'ambiguïté et rend le diagnostic de l'arthrose du genou difficile. Par conséquent, il existe un besoin considérable de développer des méthodes automatisées pour diagnostiquer l'arthrose du genou.

Motivation

Cette section traite de quelques questions ouvertes intéressantes relatives à la prédiction de la gonarthrose auxquelles nous apportons des réponses dans cette thèse.

- Absence de traitement : les thérapies existantes pour traiter l'arthrose sont limitées. La plupart des traitements actuels visent à soulager la douleur et à réduire ou à éviter le handicap dû à la destruction des os et du cartilage. Les traitements médicamenteux agissent sur les symptômes, mais pas sur la cause de cette maladie et aucun traitement ne peut stopper les changements structurels

dégénératifs responsables de son évolution. De plus, les essais cliniques de nouvelles thérapies sont complexes et très variables, avec une expression de l'arthrose différente selon les patients. Par conséquent, le diagnostic précoce de l'arthrose du genou est essentiel pour aider et permettre au patient d'adapter son mode de vie aux facteurs qui influent sur la maladie.

- Subjectivité : l'évolution des caractéristiques impliquées dans l'arthrose du genou est continue ; par conséquent, la classification en grade distinct repose souvent sur le jugement subjectif de l'opérateur. Cela introduit de la subjectivité/de l'ambiguïté et rend le diagnostic de l'arthrose du genou difficile. Par conséquent, il existe un besoin important de développer des méthodes automatisées pour aider à l'étiquetage et améliorer sa précision en identifiant les motifs et les caractéristiques dans les données liées à l'arthrose du genou qui peuvent échapper à un annotateur humain.

- Forte ressemblance : en raison de la complexité des images radiographiques et de la forte ressemblance entre les images des cas d'arthrose du genou et des sujets sains dans un stade précoce, il est difficile d'extraire des motifs pertinents pour la caractérisation de l'arthrose. Une solution proposée dans cette thèse est d'intégrer une régularisation discriminative dans le processus d'apprentissage des méthodes d'apprentissage profond supervisées et non supervisées. L'objectif était d'apprendre les propriétés les plus discriminantes et les représentations pertinentes pour détecter les signes précoces de l'arthrose du genou à partir de radiographies simples.

- Texture : dans les études cliniques sur l'arthrose, des facteurs de risque tels que l'IMC, l'âge et le sexe sont souvent utilisés pour sélectionner les personnes ayant un risque plus élevé d'évolution de l'arthrose du genou. Cependant, les effets et les interactions de ces prédicteurs ne sont pas bien connus, et les tentatives pour les utiliser pour la progression et le dépistage précoce de la gonarthrose n'ont pas été très efficaces. Les analyses de la microstructure osseuse dans l'arthrose datent de plus de 30 ans et ont montré que des modifications de l'os autour de l'articulation se produisent très tôt dans le développement de l'arthrose. Ainsi, des caractéristiques texturales invisibles à l'œil nu peuvent permettre d'extraire des informations utiles pour aider à prédire l'arthrose du genou à un stade précoce.

Le développement de modèles informatiques de prédiction peut soutenir les cliniciens dans leur choix et fournir une prédiction objective et reproductible.

Principales contributions

La principale contribution de cette thèse est l'amélioration des modèles d'apprentissage profonds (DL) pour le diagnostic précoce de l'arthrose du genou. Nous proposons d'explorer l'analyse de la texture et d'améliorer le pouvoir discriminant des modèles DL pour apprendre et extraire les caractéristiques les plus pertinentes pour les signes précoces de l'arthrose. Nous nous sommes d'abord intéressés à étendre

l'apprentissage non supervisé des caractéristiques pour faire face au problème d'une grande ressemblance entre les images de textures osseuses des patients atteints d'arthrose précoce et des sujets sains. Comme l'arthrose du genou est caractérisée par des informations de forme et de texture, nous avons optimisé le réseau neuronal convolutionnel (CNN) pour cette tâche. En plus d'appliquer la perte discriminante proposée pour améliorer la séparabilité des classes, nous avons introduit un nouveau bloc dans l'architecture CNN pour renforcer l'analyse de la texture, qui n'est pas bien prise en compte dans les CNN classiques. Les principales contributions de cette thèse sont les suivantes :

- Apprentissage d'une représentation profonde discriminante : les représentations profondes peuvent être apprises à l'aide de modèles d'apprentissage non supervisés ou supervisés. Nous proposons une nouvelle régularisation discriminative dans le processus d'apprentissage d'un réseau Auto-Encodeur (AE) non supervisé pour contraindre le réseau modifié à capturer des propriétés discriminantes qui augmentent la distance entre les classes de caractéristiques (voir chapitre 4). Nous avons également adapté la régularisation discriminative proposée à la tâche de classification supervisée multiclasse. Plus précisément, une perte discriminative a été introduite dans les représentations multiéchelles pour résoudre le problème des modèles basés sur CNN dans le cas d'une forte similarité inter-classe ou d'une forte variation intra-classe (voir chapitre 5).

- Exploitation des informations de forme et de texture : en plus de l'importance d'apprendre une représentation profonde discriminante dans la chaîne de classification, les caractéristiques extraites doivent représenter adéquatement les symptômes les plus pertinents de l'arthrose du genou. L'arthrose du genou est caractérisée par des propriétés de forme et de texture à travers l'articulation du genou. Ainsi, dans ce travail, nous introduisons un *Discriminative Shape-Texture Convolutional Neural Networks*(DST-CNN) pour mieux détecter les symptômes précoces de l'arthrose du genou. Plus précisément, nous avons amélioré le réseau CNN pour tenir compte des informations globales de forme et de texture liées aux changements micro-architecturaux osseux (voir chapitre 5).

- Effectuer une évaluation quantitative et qualitative : Une évaluation expérimentale complète a été menée sur plusieurs configurations et réglages du réseau proposé. Les hypothèses et les approches proposées ont été validées à la fois qualitativement et quantitativement. Les résultats obtenus sur deux grandes bases de données publiques, OsteoArthritis Initiative (OAI) et Multicenter OsteoArthritis Study (MOST), montrent le potentiel des méthodes proposées dans les tâches de classification binaire et multiclasse. Le modèle final proposé a dépassé la plupart des méthodes de pointe pour le diagnostic de l'arthrose du genou (voir la section Discussion au chapitre 5).

Plan de thèse et résumé des chapitres

Ce travail traite du problème du diagnostic précoce de l'arthrose du genou en optimisant les modèles d'apprentissage profonds. Cette thèse propose une revue détaillée des méthodes de pointe pour le diagnostic précoce de l'arthrose du genou à partir d'images radiographiques. Les méthodes d'apprentissage profond les plus adaptées sont présentées et discutées. Le but est de partager ce travail avec un large public de lecteurs en médecine et en informatique. Le plan du manuscrit est le suivant :

Chapitre 2 : Contexte et revue de la littérature

Ce chapitre donne un aperçu complet de l'arthrose du genou, de ses impacts et de sa sévérité et décrit la base de données des images radiographiques qui ont été utilisées. Les bases de données OsteoArthritis Initiative (OAI) et Multicenter Osteoarthritis Study (MOST) ont joué un rôle clé pour améliorer notre connaissance de la gonarthrose. Elles ont servi à élaborer des modèles prédictifs pour le diagnostic et l'évolution de la maladie, et à évaluer l'efficacité des différentes méthodes proposées. Par ailleurs, l'utilisation de ces bases de données multicentriques est essentielle pour le développement de modèles d'apprentissage profonds qui se généralisent bien, en réduisant les biais et en garantissant la précision et la fiabilité de leur prédiction. Ce chapitre fournit également une revue approfondie de la littérature des approches classiques et basées sur l'apprentissage profond existantes pour le diagnostic de l'arthrose du genou. Au sein de chaque catégorie, nous distinguons les approches selon le type de descripteur utilisé dans le cas des méthodes classiques (texture, forme, forme-texture) et le type de modèle employé dans le cas des méthodes d'apprentissage profond (AE ou CNN). Enfin, nous présentons les résultats de notre analyse sous forme de quatre tableaux, qui comparent la performance, les avantages et les inconvénients des méthodes classiques basées sur la texture (Tableau 2.2), la forme (Tableau 2.3), la forme-texture (Tableau 2.4) et les méthodes d'apprentissage profond (Tableau 2.5). Les résultats montrent que les méthodes d'apprentissage profond dépassent généralement les méthodes classiques, avec les approches forme-texture et apprentissage profond qui affichent des performances particulièrement fortes par rapport aux méthodes texture et forme seulement. Cela indique que les modèles d'apprentissage profond utilisant à la fois des informations de forme et de texture peuvent être particulièrement efficaces pour diagnostiquer avec précision l'arthrose du genou précoce.

Chapitre 3 : Introduction à l'apprentissage profond

L'apprentissage profond a fait des progrès remarquables au cours de la dernière décennie dans de nombreux domaines, notamment dans ses applications en imagerie médicale. L'apprentissage profond est un sous-domaine des méthodes d'apprentissage automatique basées sur des réseaux de neurones artificiels. Ces réseaux de neurones apprennent sans intervention humaine à effectuer des tâches complexes directement à partir des données. Pour comprendre l'idée derrière l'apprentissage profond et les méthodes

proposées dans cette thèse, nous avons dû introduire les concepts fondamentaux des réseaux de neurones profonds. Ce chapitre donne un aperçu général des réseaux de neurones à propagation avant, de l'apprentissage profond de représentation et des techniques d'apprentissage telles que la régularisation. Nous présentons notamment le réseau AutoEncodeur, qui est à la base du modèle présenté au chapitre 4. Nous présentons également des réseaux de neurones convolutionnels avancés liés au travail présenté au chapitre 5.

Chapitre 4 : Auto-encodeur régularisé discriminant pour la détection précoce de la gonarthrose

Dans ce chapitre, nous proposons une nouvelle méthode pour la détection précoce de l'arthrose du genou à partir d'images radiographiques. Nous nous concentrons principalement sur l'étape d'apprentissage des caractéristiques comme un élément essentiel du système de classification afin d'apprendre et d'extraire les caractéristiques les plus discriminantes à partir des intensités des pixels. Pour cela, nous introduisons un Auto-Encodeur Régularisé Discriminant (DRAE). Plus précisément, un terme de pénalité, appelé perte discriminative, est combiné au critère d'entraînement standard de l'Auto-Encodeur qui vise à forcer la représentation apprise à contenir des informations discriminantes. L'approche proposée effectue une série d'étapes consécutives. Tout d'abord, une méthode de segmentation semi-automatique est utilisée pour localiser l'articulation du genou. Ensuite, cinq Régions d'Intérêt (ROI) de taille 32×32 pixels (Figure 4.5) sont extraites de l'os sous le plateau tibial, représentant les parties latérale, centrale et médiale du genou. Ensuite, le modèle proposé est appliqué pour extraire des caractéristiques de haut niveau de chaque ROI. Ensuite, les caractéristiques obtenues sont utilisées pour entraîner le classifieur à différencier le cas normal (KL grade 0) et l'arthrose précoce (KL grades 1 ou 2). Une fois le classifieur entraîné, nous lui fournissons l'ensemble de données de test pour prédire leurs étiquettes correspondantes. Enfin, nous évaluons qualitativement et quantitativement les résultats obtenus. Ces étapes ont été résumées dans le schéma illustré à la figure 4.1. Des résultats expérimentaux ont été réalisés sur la base de données multicentrique OAI, qui indique que la méthode est robuste face aux artefacts et aux paramètres d'acquisition des données. Les résultats obtenus montrent que les régions les plus discriminantes se situaient dans le compartiment médial du genou, ce qui a également été montré dans d'autres études [Woloszynski et al., 2010, Janvier et al., 2017]. Cette étude prouve qu'il est encore possible d'améliorer les Auto-Encodeurs classiques pour une meilleure extraction des motifs discriminants liés à l'arthrose.

Chapitre 5 : Discriminative Shape-Texture Convolutional Neural Networks

Les approches de Deep Learning basées sur les réseaux de neurones convolutionnels (CNN) ont montré des résultats prometteurs dans la détection de l'arthrose du genou (OA) (voir section 2.3.2). Cependant, le problème de la détection précoce de l'OA du genou à partir de radiographies simples reste une tâche difficile. Cela est dû à la grande similitude entre les images OA et non-OA au stade précoce et à la

nature de l'architecture des CNN qui néglige les informations de texture liées aux changements de la microarchitecture osseuse dans leurs couches de classification. Plus précisément, plusieurs études [Wen et al., 2016, Cai et al., 2018, Cheng et al., 2018] ont montré que dans le cas de fortes similarités inter-classes ou de fortes variations intra-classes, et en utilisant seulement la perte d'entropie croisée softmax, les caractéristiques apprises avec les CNN traditionnels de la même classe sont souvent dispersées, et celles apprises de différentes classes sont chevauchées. De plus, les modèles CNN classiques conduisent à extraire des corrélations complexes dans les couches supérieures, correspondant aux informations de forme globale et négligeant les détails fins de l'image correspondant aux informations de texture [Cimpoi et al., 2015, Andrearczyk and Whelan, 2016]. Par ailleurs, d'autres études [Zeiler and Fergus, 2014, Springenberg et al., 2014] ont montré que l'architecture des CNN classiques tend à augmenter le niveau d'abstraction de la représentation avec la profondeur. Les premières couches des CNN sont conçues pour apprendre des caractéristiques de bas niveau, telles que les bords et les courbes, qui caractérisent les informations de texture, tandis que les couches plus profondes sont entraînées à capturer des motifs plus complexes et de haut niveau, tels que les informations de forme globale. Par conséquent, à la suite de convolutions successives, de fonctions d'activation et d'opérations de regroupement, les détails fins de l'image liés à la texture disparaissent dans les couches supérieures du réseau. Comme mentionné ci-dessus, l'OA du genou est représentée par des propriétés de forme et de texture sur l'ensemble de l'articulation du genou. De plus, aux stades précoces de la maladie, les radiographies sont souvent très similaires. Il est donc important que le modèle proposé prenne en compte ces problèmes. Le présent travail vise à améliorer le diagnostic automatique précoce de l'OA du genou en utilisant un réseau neuronal convolutionnel. Tout d'abord, inspirés par des recherches antérieures sur les CNN de texture [Cimpoi et al., 2015, Andrearczyk and Whelan, 2016] et notre régularisation discriminative récemment proposée [Nasser et al., 2020], nous introduisons un modèle CNN profond appelé DCNN (Discriminative Convolutional Neural Network) [Nasser et al., 2022] pour prendre en compte à la fois les changements de forme et de texture et maximiser la séparabilité entre les sujets OA et non-OA. Ensuite, nous proposons une extension du DCNN à un réseau DST-CNN (Discriminative Shape-Texture Convolutional Neural Networks) [Nasser et al., 2023] pour améliorer les résultats de classification en (i) améliorant la qualité des informations de texture avant de les combiner avec la forme globale du genou et (ii) incorporant une nouvelle perte discriminante pour améliorer la séparabilité des classes d'OA du genou (KL-0, KL-1 et KL-2) dans un cadre de classification multi-classes. Le potentiel de nos réseaux proposés a été évalué pour des tâches de classification binaire et multi-classes.Les principales contributions de cette étude sont les suivantes :

- Développement d'un nouveau réseau basé sur les CNN, capable d'apprendre des représentations hautement discriminantes et de fusionner à la fois les caractéristiques de forme et de texture, pour offrir des résultats de classification pertinents.
- La robustesse des réseaux proposés est évaluée face aux artefacts et aux paramètres d'acquisition

des données en réalisant des expériences sur deux grandes bases de données publiques : la base de données Multi-center Osteoarthritis Study (MOST) utilisée pour l'entraînement et la base de données OsteoArthritis Initiative (OAI) utilisée pour la validation et les tests.

- Une étude d'ablation est réalisée pour évaluer l'impact de chaque composant du réseau proposé sur le processus d'apprentissage.
- La transparence dans le processus de décision est assurée grâce à des cartes d'activation montrant la contribution des différentes zones du genou à la décision finale.

L'analyse expérimentale a montré une amélioration des performances de classification en utilisant le réseau proposé, par rapport aux réseaux de l'état de l'art. Grâce à la technique de visualisation Grad-CAM, nous avons démontré que dans la plupart des cas, le modèle apprend des caractéristiques intéressantes pertinentes pour le processus de décision. De plus, l'étude d'ablation a montré l'utilité et l'efficacité de chaque composant du réseau proposé. À notre connaissance, il s'agit du premier travail basé sur un réseau neuronal profond qui combine les caractéristiques de forme et de texture pour le diagnostic automatique précoce de l'arthrose du genou. L'approche proposée peut être facilement intégrée à divers modèles CNN et peut être utilisée dans différentes tâches d'imagerie médicale avec d'autres types de données.

Conclusion

Aujourd'hui, le diagnostic précoce de l'arthrose du genou est essentiel pour proposer un traitement efficace avant de faire face à une pathologie sévère et irréversible et pour soutenir et permettre aux patients de prendre en compte les facteurs de mode de vie qui influencent la maladie. Au cours de la dernière décennie, le deep learning a suscité beaucoup d'attention de la part de la communauté scientifique et a connu un grand succès dans de nombreuses applications d'imagerie médicale. Cependant, malgré ces succès, le diagnostic précoce de l'arthrose du genou à partir de radiographies simples est resté une tâche très difficile. Le besoin d'améliorer les modèles de deep learning pour mieux prédire les symptômes précoces de l'arthrose a été l'une des principales motivations qui ont conduit à l'initiation de cette thèse. Cette section résume la thèse, met en évidence les contributions proposées et discute de plusieurs pistes de recherche pour les travaux futurs.

Dans cette thèse, nous avons travaillé à améliorer la séparabilité des classes entre les cas Normaux (KL-0), Douteux (KL-1) et OA Minime (KL-2) en classification binaire et multi-classes. Pour cela, nous avons étudié et abordé le problème de l'utilisation du deep learning pour prédire automatiquement les symptômes précoces et les changements osseux. Le matériel du chapitre 4 a proposé un nouveau modèle d'apprentissage de représentation, appelé Auto-Encodeur Régularisé Discriminant (DRAE), pour découvrir automatiquement des représentations latentes intéressantes grâce à un apprentissage non supervisé classique et une nouvelle régularisation discriminative. La régularisation discriminative fonctionne

en forçant le réseau à capturer non seulement les caractéristiques importantes présentes dans les données, mais aussi les plus discriminantes qui minimisent la variation intra-classe et maximisent la distance inter-classe. Cette première contribution est une étape dans la direction de rendre les modèles de DL utiles dans le cas d'une forte similitude entre les classes, que nous continuons à explorer dans les chapitres suivants. Comme le DRAE est un réseau entièrement connecté, il ne parvient pas à capturer les motifs complexes dans les images car il ne tient pas compte des informations de voisinage. De plus, le nombre de poids augmente rapidement avec la taille de l'image, devenant ingérable. Alors que les DRAE se concentrent sur une petite région sous le plateau tibial, l'OA peut se produire dans n'importe quelle zone de l'articulation du genou. Dans la première partie du chapitre 5, nous avons décrit un réseau neuronal convolutionnel discriminant pour mieux capturer les motifs discriminants de toute la zone de l'articulation du genou. La méthode proposée améliore la qualité de la prédiction précoce de l'OA en incorporant une perte discriminative à la fonction objective standard des CNN pour renforcer la séparation entre les caractéristiques des patients OA et non-OA, et une stratégie de concaténation de caractéristiques multi-échelles pour améliorer la représentation des propriétés fines dans la couche de classification supérieure. Dans le même chapitre, nous abordons également le problème d'améliorer l'analyse de texture en utilisant l'idée principale du CNN discriminant. Nous introduisons un réseau DST-CNN (Discriminative Shape-Texture Convolutional Neural Networks) qui étend l'approche précédente en ajoutant un nouveau bloc pour extraire les informations de texture en calculant les corrélations entre les cartes de caractéristiques dans plusieurs couches intermédiaires. Nous adaptons également la perte discriminative pour qu'elle corresponde à la tâche de classification multi-classes. En utilisant le DST-CNN, nous pouvons mieux caractériser les informations de forme et de texture nécessaires pour détecter les symptômes précoces et les changements osseux liés à l'OA.

Perspectives

Des travaux futurs qui peuvent être envisagés pour prolonger cette recherche et développer davantage le domaine de l'apprentissage de caractéristiques discriminantes et de l'analyse de texture en utilisant le deep learning sont les suivants :

- Améliorer le modèle Auto-Encodeur Régularisé Discriminant (DRAE)
  - Étendre l'architecture du modèle : La nature entièrement connectée du DRAE limite sa capacité à capturer les motifs dans les données de pixels qui peuvent être influencés par les informations de voisinage. Des recherches futures pourraient étudier l'incorporation de couches convolutionnelles ou d'autres techniques pour exploiter les relations spatiales au sein de la zone de l'articulation du genou. Cela aiderait le modèle à mieux comprendre le contexte local et à améliorer potentiellement ses performances dans la détection des caractéristiques liées à l'OA.

- Perte de reconstruction hybride : Une autre direction possible pour améliorer le modèle DRAE est d'incorporer une perte de reconstruction hybride qui combine les capacités de reconstruction des bords avec la préservation de la texture. Cette amélioration peut permettre au modèle auto-encodeur de capturer à la fois les informations structurelles et texturales, conduisant à une meilleure précision de classification et une détection plus robuste des indicateurs précoces de l'OA.

- Améliorer le réseau DST-CNN (Discriminative Shape-Texture Convolutional Neural Networks)
  - Explorer des fonctions de perte alternatives : La fonction de perte discriminative proposée a été efficace pour renforcer la séparation entre les caractéristiques OA et non-OA. Cependant, il y a de la place pour explorer des fonctions de perte ou des techniques de régularisation alternatives qui peuvent augmenter davantage le pouvoir discriminant du modèle. Des fonctions de perte adaptatives ou dynamiques qui peuvent ajuster de manière adaptative l'accent sur différentes classes ou caractéristiques pourraient être étudiées pour améliorer les performances de classification.
  - Améliorer l'analyse de texture : L'analyse de texture est cruciale pour détecter les symptômes précoces et les changements osseux liés à l'OA. Nous pouvons améliorer le module d'analyse de texture du DST-CNN en incorporant des techniques plus avancées inspirées du transfert de style, de la synthèse de texture ou du remplissage de texture. Ces techniques peuvent aider à préserver ou à restaurer les détails fins et les textures des images d'entrée. Ces améliorations pourraient rendre le DST-CNN plus précis et fiable pour l'évaluation de la gravité de l'OA.

- Autres directions futures potentielles
  - Incorporer des données multimodales : Dans cette thèse, nous nous sommes principalement concentrés sur l'utilisation des données d'image radiographique pour la classification de l'OA. Cependant, l'OA est une condition complexe qui peut impliquer diverses modalités, telles que des rapports textuels, des données démographiques des patients ou des données génétiques. Des travaux futurs peuvent explorer l'intégration de sources de données multimodales pour améliorer la précision et la robustesse des modèles de classification. Cela pourrait impliquer le développement de techniques de fusion ou l'incorporation du mécanisme d'attention Transformer pour exploiter les informations provenant de différentes modalités.
  - Adaptation de domaine : L'application de modèles de deep learning entraînés sur des jeux de données d'images naturelles comme ImageNet à un autre jeu de données médicales comme OAI et MOST avec des caractéristiques différentes entraîne souvent une dégradation des performances due au décalage de domaine. L'étude de méthodes d'adaptation de domaine peut aider à combler le fossé entre les différents jeux de données et à améliorer la généralisabilité

des modèles de deep learning proposés pour la classification de l'OA. Cependant, l'OA est une condition complexe qui peut impliquer diverses modalités, telles que des rapports textuels, des données démographiques des patients ou des données génétiques. Des travaux futurs peuvent explorer l'intégration de sources de données multimodales pour améliorer la précision et la robustesse des modèles de classification. Cela pourrait impliquer le développement de techniques de fusion ou l'incorporation du mécanisme d'attention Transformer pour exploiter les informations provenant de différentes modalités.

- Adaptation de domaine : L'application de modèles de deep learning entraînés sur des jeux de données d'images naturelles comme ImageNet à un autre jeu de données médicales comme OAI et MOST avec des caractéristiques différentes entraîne souvent une dégradation des performances due au décalage de domaine. L'étude de méthodes d'adaptation de domaine peut aider à combler le fossé entre les différents jeux de données et à améliorer la généralisabilité des modèles de deep learning proposés pour la classification de l'OA. Cela comprend l'exploration de techniques telles que l'adaptation de domaine non supervisée, l'extraction de caractéristiques spécifiques au domaine et les stratégies d'augmentation de données spécifiques aux domaines cibles.

Liste des publications :

Les publications suivantes sont issues des travaux réalisés dans le cadre de cette thèse :

- Articles de revue
  - Yassine Nasser, Rachid Jennane, Aladine Chetouani, Eric Lespessailles, Mohammed El Hassouni. "Discriminative Regularized Autoencoder for Early Detection of Knee Osteoarthritis: Data from the Osteoarthritis Initiative", IEEE TRANSACTIONS ON MEDICAL IMAGING, vol. 39, no. 9, pp. 2976-2984, Sept. 2020. [Nasser et al., 2020]
  - Yassine Nasser, Mohammed El Hassouni, Didier Hans, Rachid Jennane. "A Discriminative Shape-Texture Convolutional Neural Network for Early Diagnosis of Knee Osteoarthritis from Radiographic Images", Physical and Engineering Sciences in Medicine, 827–837 (2023). [Nasser et al., 2023]
- Articles de conférence1
  - Nasser, Yassine, Mohammed El Hassouni, and Rachid Jennane. "Deep Discriminative Neural Network for Predicting Knee Osteoarthritis at an Early Stage." International Workshop on PRedictive Intelligence In MEdicine MICCAI. Springer, Cham, 2022. [Nasser et al., 2022]

- Yassine Nasser, Mohammed El Hassouni, Abdelbasset Brahim, Hechmi Toumi, Mohamed Hedi Bedoui, Eric Lespessailles, Rachid Jennane, “Diagnosis of osteoporotic disease from bone radiographic images with Sparse Stacked Autoencoder and SVM classifier”, 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP’2017), 2017, pp. 1-5. [Nasser et al., 2017]

- Autres

  - Yassine Nasser, Abdessamad Tafraouti, Mohammed El Hassouni, Hechmi Toumi, Mohamed Hedi Bedoui, Eric Lespessailles, Rachid Jennane, “Diagnosis of osteoporosis using deep autoencoder with SVM”, 8th WORKSHOP AMINA “Medical Applications of Computer Science: New Approaches”, November 17th to 19th 2016 in Monastir, Tunisia. [Nasser et al., 2016]

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

Acc Accuracy

AE Auto-Encoder

AI Artificial Intelligence

AUC Area Under the Curve

BMI Body Mass Index

CAD Computer-Aided Diagnosis

CAE Contractive AutoEncoder

CE Cross-Entropy

CNN Convolutional Neural Network

CR computed radiography

DAE Denoising AutoEncoder

DCNN Discriminative Convolutional Neural Network

DST-CNN Discriminative Shape-Texture Convolutional Neural Network

DL Deep Learning

DRAE Discriminative Regularized Auto-Encoder

DST-DNet Discriminative Shape-Texture DenseNet

F1 F1-score

GAP Global Average Pooling

GMD Gram Matrix Descriptor

Grad-CAM Gradient-weighted Class Activation Mapping

JSN Joint Space Narrowing

KNN K-Nearest Neighbours

KL Kellgren and Lawrence

KLD Kullback-Leibler Divergence

LDA Linear Discriminant Analysis

MAE Mean Absolute Error

ML Machine Learning

MLR Multivariate Linear Regression

MLP Multi-Layer Perceptron

MOST Multicenter Osteoarthritis Study

MSE Mean Square Error

OA OsteoArthritis

OAI Osteoarthritis Initiative

PA PosteroAnterior

Pr Precision

Re Recall

RMSE Root Mean Square Error

ROI Regions Of Interest

SAM Separable Adaptive Max-Pooling

SAE Sparse AutoEncoder

SMC Softmax Classifier

SoA state-of-the-art

SSL Semi-Supervised Learning

TBT Trabecular Bone Texture

t-SNE t-distributed Stochastic Neighbor Embedding

1

Introduction

Contents

1.1 Context and Challenge

Artificial intelligence (AI) has made exponential advances over the past decade. This progress is evidenced by its impressive success in many applications in different fields, including healthcare, business, education, autonomous vehicles, and social media. Computer vision is an interesting domain of AI that deals with the science of making computers or machines visually enabled, i.e., they can analyze and understand an image automatically. Computer vision is a broad research field that involves extracting information and understanding images using computer algorithms. This domain encompasses various problems, including detection, segmentation, recognition, motion estimation, and image restoration. Although computer vision aims to address the challenge of understanding an image, Machine Learning (ML) focuses on understanding the vast amount of available data.

1.2 Motivation for the thesis

This section addresses some interesting open questions regarding knee OA prediction that we address in this thesis.

No treatment **:** Therapies available for treating osteoarthritis are limited. Most current treatments are designed to relieve pain and reduce or prevent disability caused by bone and cartilage degeneration. Drug therapies target the symptoms but not the cause of this disease, and no treatment inhibits the degenerative structural changes responsible for its progression. Furthermore, clinical testing of new therapies is complicated by the highly variable manner in which OA manifests in individual patients. Hence, early diagnosis of knee OA is essential in supporting and enabling the patient to address lifestyle factors that affect the disease.

Subjectivity **:** The evolution of the features involved in knee OA is continuous; therefore, the classification into distinctive grades is often left to the subjective opinion of the operator. This introduces subjectivity/ambiguity and makes the diagnosis of knee OA challenging. Consequently, there is a considerable need to develop automated methods to assist in the labeling process and improve labeling accuracy by identifying patterns and features in the data related to knee OA that may not be visible to a human annotator.

High similarity **:** Due to the complex nature of X-ray images and the high similarity between the images of knee OA cases and healthy subjects, extracting meaningful patterns relevant to OA characterization is challenging. One solution proposed in this thesis is to incorporate discriminative regularization into the learning process of both unsupervised and supervised deep learning methods. The objective was to learn the most discriminating properties and the relevant representations to detect the early signs of knee OA from plain radiographs.

Texture **:** In OA clinical trials, risk factors such as BMI, age, and sex are commonly used to select

individuals at greater risk of knee OA progression [Culvenor et al., 2015]. Unfortunately, the effects and interactions of these predictors are not fully understood, and efforts to use them for knee OA progression and early OA detection have not been highly successful. Analyses of the bone microarchitecture in OA date back more than 30 years and have provided clear indications that changes in the periarticular bone occur very early in OA development [Lynch et al., 1991a, Goldring, 2009]. Thus, textural features not visible to the naked eye can enable the extraction of relevant information to help predict knee OA at an early stage.

Developing computer-based prediction models can support clinicians in their decisions and provide an objective and reproducible prediction.

1.3 Thesis summary and main contributions

The main contribution of this thesis is the improvement of Deep Learning (DL) models for the early diagnosis of knee OA. We propose to investigate texture analysis and improve the discriminative power of DL models to learn and extract the features most relevant to early signs of OA. We first focused on extending unsupervised feature learning to deal with the problem of a high degree of similarity between the bone texture images of patients with early OA and healthy subjects. Because knee OA is characterized by both shape and texture information, we improved the Convolutional Neural Network (CNN) for this task. In addition to applying the proposed discriminant loss to improve class separability, we introduced a new block into the CNN architecture to enhance texture analysis, which is not well considered in classical CNNs. The main contributions of this thesis are as follows :

Learning deep discriminative representation : Deep representations can be learned using unsupervised or supervised learning models. We propose a new discriminative regularization in the learning process of an unsupervised Auto-Encoder (AE) network to force the modified network to capture discriminative properties that maximize the distance between feature classes (see Chapter 4). We also improved the proposed discriminative regularization to fit the supervised multiclass classification task. More specifically, a discriminative loss was introduced into the multi-scale representations to address the problem of CNN-based models in the case of high inter-class similarities or high intra-class variations (see Chapter 5).

Exploiting both shape and texture information : In addition to the importance of learning a deep discriminative representation in the classification pipeline, the extracted features must adequately represent the most relevant symptoms of knee OA. Knee OA is characterized by shape and texture properties across the knee joint. Thus, in this work, we introduce a Discriminative Shape-Texture CNN (DST-CNN) to better detect early knee OA symptoms. Specifically, we improved the CNN network to consider the overall shape and texture information related to bone microarchitectural changes (see Chapter 5).

Conducting quantitative and qualitative evaluation : A comprehensive experimental evaluation was conducted on several configurations and settings of the proposed network. The proposed hypotheses and approaches are verified both qualitatively and quantitatively. The results obtained on two large public databases, OsteoArthritis Initiative (OAI) and Multicenter OsteoArthritis Study (MOST), demonstrate the potential of the proposed methods in both binary and multiclass classification tasks. The final proposed model outperformed most state-of-the-art knee OA diagnosis methods (see the Discussion section in Chapter 5).

1.4 Thesis outline

This work addresses the problem of early diagnosis of knee OA by improving deep-learning models. This thesis provides a detailed review of state-of-the-art approaches for early diagnosis of knee OA using X-ray images. The most relevant DL methods are presented and discussed. The goal is to share this work with a wide range of medical and computer science readers. The outline of the manuscript is as follows :

- **Context and Literature Review :**

  Chapter 2 gives a comprehensive overview of knee OA, its impacts, and severity and describes the database of the X-ray images that were used. This chapter also provides a thorough literature review of the existing classical and deep learning-based approaches for diagnosing knee OA.

- **Deep Learning Overview**

  Chapter 3 presents an overview of existing deep-learning networks and the motivation behind their use. This chapter also introduces the recent evolution and extensions of CNNs to better capture the texture and learn more discriminative features.

- **Discriminative Regularized Auto-Encoder**

  Chapter 4 introduces a Discriminative Regularized Auto-Encoder (DRAE) for the early diagnosis of knee OA using X-ray images. This chapter focuses on analyzing the texture to detect the microarchitectural bone changes caused by OA. Due to the high similarity between the X-ray images of OA cases and healthy subjects, we propose training the DRAE network to capture discriminative properties that minimize the intra-class distance and maximize the inter-class distances. To achieve this, an additional term called discriminative penalty is added to the initial reconstruction cost function of the standard AE. This leads the proposed network to map the input image to an encoding space that maximizes the distance between the feature classes. The effectiveness of this method in producing more discriminative features was demonstrated by conducting quantitative and qualitative experiments on a public OAI database.

- **Discriminative Shape-Texture Convolutional Neural Network (DST-CNN)**

  Chapter 5 focuses on CNN-based models to exploit both texture and shape information to better detect early signs of knee OA. Classical CNNs with cross-entropy loss do not perform efficiently in separating data with high inter-class similarities or intra-class variations. In addition, the CNN architecture extracts complex correlations and neglects fine image details related to the texture in their classification layers. In this chapter, we propose a new CNN-based network called DST-CNN to overcome such issues. The proposed DST-CNN extends the discriminative regularization proposed in Chapter 4 to fit the multiclass classification task between healthy, doubtful, and mild OA. In the DST-CNN, texture and shape information was considered by computing the correlations between feature maps in several intermediate layers and combining them with the global average pooling in the top layers. We evaluated the DST-CNN model by using two large public databases (MOST and OAI). A comparison with state-of-the-art DL was also provided. Various experiments, including ablation studies, have been conducted to demonstrate the usefulness and effectiveness of each network component. Grad-CAM visualization was used to visualize attention maps and enhance the interpretability of the models.

- **Conclusion**

  Chapter 6 summarizes the contributions of this work and discusses several interesting directions for future research. A complete list of publications is also included at the end of this chapter.

2

Context and Literature Review

Contents

2.1 Knee Osteoarthritis

2.1.1 Osteoarthritis and its burdens

Osteoarthritis (OA) is the most common joint disease. It is defined as a degenerative joint disease caused by irreversible deterioration of the joint cartilage located at the end of the bone and is generally characterized by stiffness, swelling, pain, and a grating sensation on movement [Cross et al., 2014]. OA usually begins around the age of 50 and affects women more than men. Approximately 21 million American adults have physician-diagnosed OA, a diagnosis usually based on a combination of joint symptoms and radiographic changes. However, many patients have undiagnosed or subclinical diseases. The prevalence of OA in the population is difficult to determine because the degree of radiological changes in symptomatic individuals varies greatly, and many individuals with radiographic evidence of OA have no symptoms. The World Health Organization estimates that 9.6% of men and 18.0% of women aged over 60 years have symptomatic osteoarthritis, and it is estimated that 80% of those with OA will have limitations in movement, and 25% will not be able to perform the major daily activities of their lives. It has been reported that degenerative joint disease disorders, such as OA, will affect at least 130 million individuals worldwide by 2050 [Lim and Lau, 2011]. Therefore, this disease places a huge burden on healthcare services (accounting for 1–2.5% of the gross national product in Western countries), and the cost of OA for these services is expected to double by 2030. Worse still, no treatment can inhibit the degenerative structural changes responsible for the progression of knee OA. Most current treatments are designed only to relieve pain and reduce or prevent disability caused by bone and cartilage degeneration. Moreover, clinical testing of new therapies is complex in a highly variable manner, with arthritis manifesting in individual patients.

2.1.2 Knee Osteoarthritis causes and severity

Knee OA is caused by a breakdown of knee articular cartilage and bone micro-architecture changes [Heidari, 2011, Bijlsma et al., 2011]. The disease is well-recognized as the leading cause of mobility impairment in older adults and is now recognized as an independent risk factor for increased mortality. The exact causes of knee OA are not fully understood, but there are several factors that can contribute to its development:

- **Age:** As people get older, the risk of developing knee osteoarthritis increases. This is because the cartilage in the knee joint naturally wears down over time.
- **Gender:** Women are more likely to develop knee OA than men.
- **Genetics:** There is evidence to suggest that genetics play a role in the development of knee OA. People with a family history of the condition are more likely to develop it themselves.

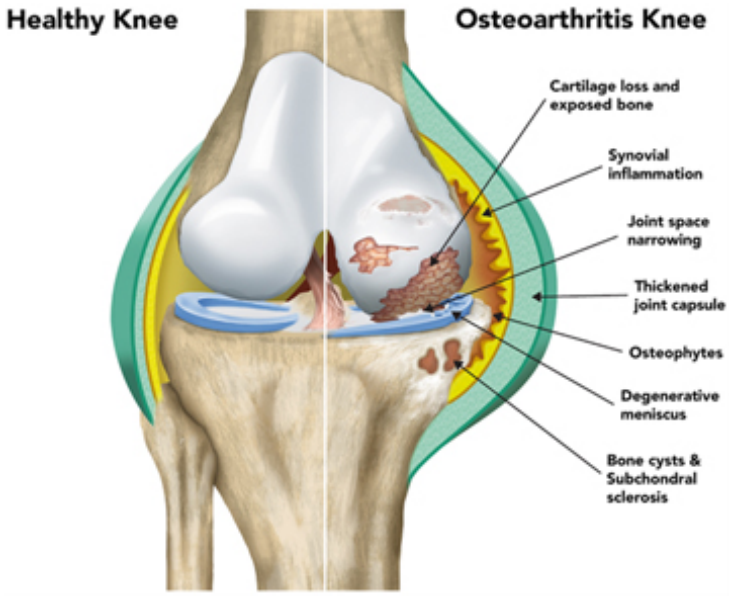


Figure 2.1: Schematic illustration showing the key pathological features of knee osteoarthritis. The left side shows the normal joint, and the right side shows the diseased joint [Cibrián Uhalte et al., 2017].

- **Obesity:** Excess weight puts extra pressure on the knee joint, which can accelerate the degeneration of cartilage.
- **Joint injuries:** Previous injuries to the knee joint, such as a torn ligament or meniscus, can increase the risk of developing knee OA.
- **Repetitive stress:** Jobs or activities that require repetitive stress on the knee joint, such as construction work or running, can increase the risk of developing knee OA.

The severity of knee OA can vary widely, from mild to severe. Common symptoms include pain, stiffness, swelling, and a reduced range of motion in the knee joint. Knee OA is typically assessed through imaging tests, such as Radiography (X-Ray), Magnetic Resonance imaging (MRI) or Computed Tomography (CT-scan), which can depict most of the hallmarks of knee OA [Lee et al., 2021] (see Figure 2.2). Radiography is considered the gold standard for knee OA assessment because of its wide accessibility, cost efficiency, and safety. The knee OA hallmarks include Joint Space Narrowing (JSN), cartilage loss, bone spurs (osteophytes), subchondral sclerosis and other changes in the joint, illustrated in Figure 2.1. Based on these features, Kellgren and Lawrence (KL) [Kellgren and Lawrence, 1957] divided knee OA severity into five stages (see Figure 2.3):

- **Normal (Grade 0):** No joint space narrowing or osteophytes are present. The knee joint appears normal on X-ray.
- **Doubtful (Grade 1):**: There is minimal joint space narrowing and/or a small osteophyte is present. This stage is often asymptomatic, and the patient may not have any symptoms.

Figure 2.2: Different medical imaging modalities for diagnosing knee osteoarthritis: (a) X-ray imaging, (b) magnetic resonance imaging (MRI), and (c) computed tomography (CT) scans.

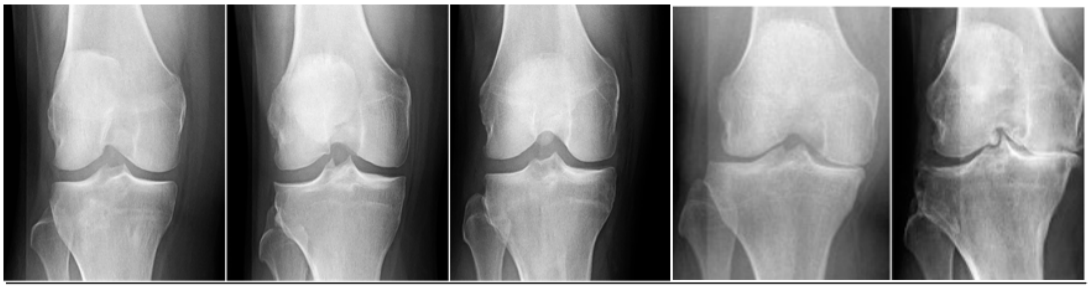


Figure 2.3: The Kellgren-Lawrence system for grading the severity of knee osteoarthritis based on radiographic features. The system has five grades from 0 (normal) to 4 (severe), with increasing levels of joint space narrowing, osteophyte formation, subchondral sclerosis and bone deformity.

- **Mild (KL Grade 2):** There is mild joint space narrowing and/or small osteophytes are present. Patients may experience mild pain and stiffness, especially after prolonged periods of activity or after sitting for extended periods of time.

- **Moderate (Grade 3):** Moderate joint space narrowing and/or moderate-sized osteophytes are present. Patients may experience more frequent pain and stiffness, especially during weight-bearing activities, and may have some limitations in their ability to perform daily activities.

- **Severe (Grade 4):** There is severe joint space narrowing and/or large osteophytes are present. Patients may experience constant pain and stiffness, even at rest, and may have significant limitations in their ability to perform daily activities. In severe cases, joint deformity may also be present.

However, the evolution of the features involved in knee OA is continuous; therefore, classification into distinct grades is often left to the subjective opinion of the annotator. This introduces subjectivity/ambiguity and makes the diagnosis of knee OA challenging. Consequently, there is a considerable need to develop automated methods to diagnose knee OA.

2.2 OsteoArthritis Database

The OsteoArthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST) databases are two valuable public resources that have been extensively studied in the field of osteoarthritis research. These databases provide researchers with access to a vast amount of data related to osteoarthritis, including clinical, radiographic, and biomarker data.

2.2.1 OsteoArthritis Initiative (OAI)

The Osteoarthritis Initiative (OAI) is a multicenter study conducted by four clinical centers and a data coordinating center [Peterfy et al., 2008]. The OAI is a longitudinal study of $4,796$ men and women aged 45-79 years, and documents the natural history of knee OA across the complete spectrum of diseases including at-risk subjects, those with early/preclinical disease, subjects with established OA, and those with end-stage disease. The OAI is sponsored by the National Institutes of Health (Part of the Department of Health and Human Services). The goal of this study was to provide resources to better understand the prevention and treatment of knee osteoarthritis. More specifically, OAI supports the development and validation of biochemical, genetic (blood and urine), and imaging biomarkers (magnetic resonance imaging and radiography) for a better understanding of the development and progression of knee OA, and thus better prevention and treatment of this debilitating disease.

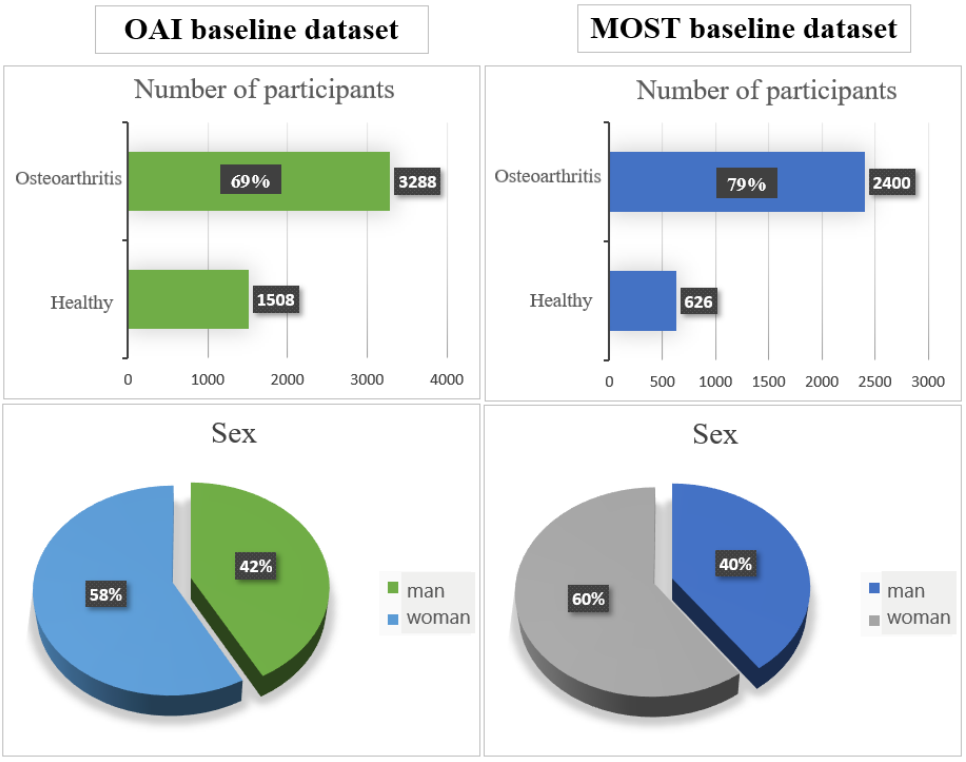


Figure 2.4: Summary statistics of the OAI and MOST datasets.

The OAI study design, measures, clinical data, and DICOM images can be downloaded online from the following website: `https://nda.nih.gov/oai`. Moreover, a variety of quantitative and semi-quantitative image evaluations are available. These data allow assessments, hypothesis generation, and testing of the development of early OA in clinically significant diseases. More than 400 research manuscripts have been generated based on these data.

2.2.2 Multicenter Osteoarthritis Study (MOST)

The Multicenter Osteoarthritis Study (MOST) is a large prospective epidemiologic cohort study, similar to the OAI study, but includes data from older subjects aged 50–79 years old [Segal et al., 2013]. All 3,026 study participants had symptomatic knee OA at baseline or were at high risk of developing the disease. MOST is a collaborative effort developed by investigators at four core sites: two clinical centers, a data coordinating center, and an analysis center. This study was also sponsored by the National Institutes of Health/National Institute on Aging, part of the Department of Health & Human Services). The main aim of MOST is to identify the risk factors for incident symptomatic knee OA and progressive knee OA, which helps us better understand how to prevent and treat this disease.

Both databases have been instrumental in advancing our understanding of osteoarthritis. They have been used to develop predictive models for disease diagnosis and progression, and to test the effectiveness

of different proposed approaches. Moreover, the use of such multicenter databases is crucial for the development of deep learning models that generalize well, minimizing bias and ensuring the accuracy and reliability of their predictions.

Table 2.1: THE CATEGORIZATION OF DIFFERENT KNEE OA CLASSIFICATION APPROACHES USING X-RAY IMAGES.

Categorize		Knee OA Classification Approaches
Handcrafted feature based	**Texture**	[Woloszynski et al., 2012] Signature Dissimilarity Measure
		[Hirvasniemi et al., 2014] Laplacian and LBP-based method
		[Janvier et al., 2017] Fractal Analysis
		[Hladůvka et al., 2017] Fractal Analysis and Shannon Entropy
		[Riad et al., 2018] Complex Wavelet Decomposition
		[Brahim et al., 2019] Circular Fourier Filter and ICA
	Shape	[Haverkamp et al., 2011] Statistical Shape Mode
		[Minciullo and Cootes, 2016] PCA and Random Forest
		[Shamir et al., 2008] Image descriptors and transforms
	Shape-Texture	[Thomson et al., 2015] Fractal Signature and Statistical Shape Model
		[Minciullo et al., 2017] Hough Forest and Combined Appearance Models
Deep Learning based	**CNNs**	[Antony et al., 2017] CNN
		[Tiulpin et al., 2018] Siamese CNN
		[Chen et al., 2019] CNN with a novel ordinal loss
		[Norman et al., 2019] Densely Connected CNN
		[Nguyen et al., 2020] CNN with SSL and Manifold Regularization
	CNN+ViT	[Konwer et al., 2022] Temporal Context Matters

2.3 State of the Art Knee Osteoarthritis Detection Approaches

2.3.1 Classic approaches

In recent decades, several methods have been developed for knee OA diagnosis using X-ray images. As a major step, features are extracted to capture changes related to the different stages of knee OA. Most current methods are based on exploiting low-level hand-crafted features to detect changes due to OA, such as texture [Woloszynski et al., 2012, Hirvasniemi et al., 2014, Janvier et al., 2017, Riad et al., 2018, Brahim et al., 2019], shape [Haverkamp et al., 2011, Minciullo and Cootes, 2016], and both shape and texture [Shamir et al., 2008, Thomson et al., 2015, Minciullo et al., 2017]. Table 2.1 enumerates several of these studies.

2.3.1.A Texture-based approaches

As Table 2.1 illustrates, most handcrafted feature approaches are based on texture. In [Woloszynski et al., 2012] Woloszynski *et al.* developed a rotation-invariant trabecular bone texture classification approach using a Signature Dissimilarity Measure (SDM) that quantifies roughness, degree of anisotropy, and direction of anisotropy of trabecular bone textures. This study used 203 radiographic images : 68 images with OA and 135 images without OA. The proposed method consisted of three consecutive stages. First, an automated region selection method was used to determine the trabecular bone Region Of Interest (ROI) on the radiographs. Two ROIs of size 112×112 were extracted on the subchondral bone immediately under the cortical plate of the medial and lateral tibial compartments, respectively (Figure 2.5). Next, three trabecular bone texture parameters corresponding to roughness, degree of anisotropy, and direction of anisotropy were calculated using the signature dissimilarity measure method. Finally, the authors evaluate the predictive abilities of the texture parameters using a binary logistic regression

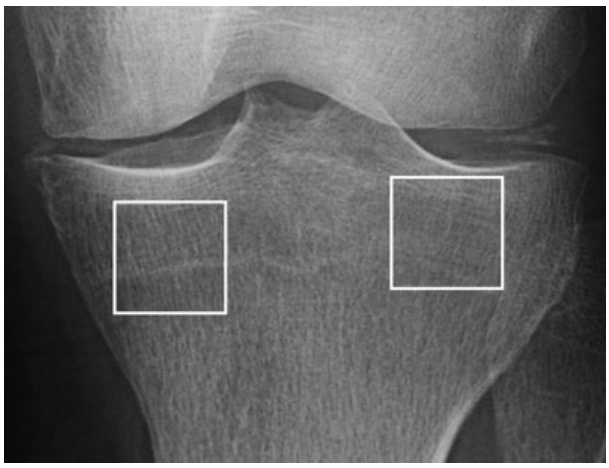


Figure 2.5: The lateral and medial trabecular bone ROIs used in [Woloszynski et al., 2012]

model. As a result, this study confirmed that medial tibial trabecular bone texture is predictive of loss of medial joint space in knees with OA. An important limitation of this study is using a small database with two different radiographic protocols.

In [Hirvasniemi et al., 2014], Hirvasniemi *et al.* propose another method to quantify the differences in bone texture between subjects with different stages of knee OA and controls using plain radiographs. This study used 203 X-ray images classified according to the Kellgrene Lawrence (KL) grading scale (KL0 : n=110, KL1 : n=28, KL2 : n=27, KL3 : n=31, KL4 : n=7). From each image, six rectangle-shaped ROIs were extracted from the tibia and femur, and one elliptical-shaped ROI from the soft tissue beside the joint (see Figure 2.6). Two ROIs (size : 70×30 pixels) were placed into the subchondral bone plate in the center of the medial and lateral condyles of the tibia, two ROIs (70×70 pixels) immediately below the subchondral bone plate in the subchondral trabecular bone in the tibia, and two ROIs in the medial (70×70 pixels) and lateral condyles of the femur (70×30 pixels). Their proposed method estimates bone density directly from the grayscale values of the unprocessed ROIs, and calculates the structure-related parameters from Laplacian- and local binary patterns (LBP) ROIs. After extracting the bone density and structure-related parameters, statistical analysis was performed. The results indicate that changes in bone texture in knee OA can be quantitatively evaluated using plain radiographs. The most significant changes were observed in the medial subchondral bone plate and trabecular bone in the proximal tibia and medial femur. Structural analysis of bone is more reproducible than direct evaluation of grayscale values, and it is, therefore, better suited for quantitative analysis.

Janvier *et al.* [Janvier et al., 2017] proposed another method to predict the incidence of radiographic knee osteoarthritis using subchondral tibial bone texture. The objective of this study was to evaluate whether Trabecular Bone Texture (TBT) parameters measured on X-ray images could predict the onset

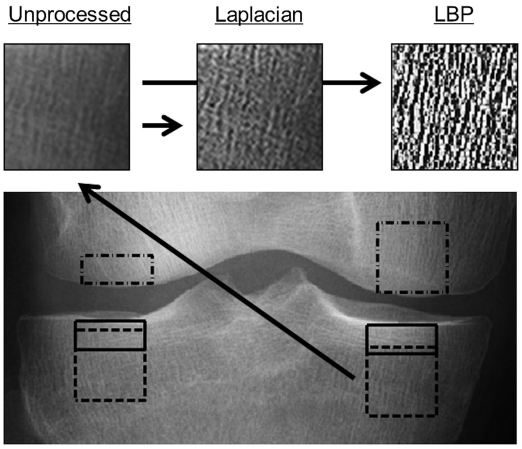


Figure 2.6: Schematic describing the location of the ROIs used in [Hirvasniemi et al., 2014]. The black-colored rectangles with continuous lines represent the subchondral bone plate ROIs. The black squares with dashed lines represent the tibial subchondral trabecular bone ROIs. The femur ROIs are shown with black rectangles with a dash-dotted line.

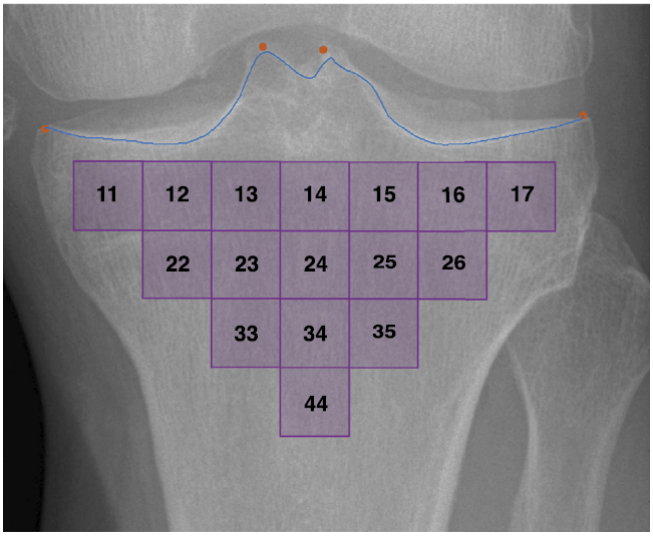


Figure 2.7: The sixteen ROIs studied by Janvier *et al.* [Janvier et al., 2017]

of radiographic knee OA. In this study, we used 344 images from the OAI database=. Sixteen ROIs were extracted from each image, as shown in figure 2.7. The authors used fractal analysis to analyze the trabecular bone texture. Their proposed method computes fractal parameter H to characterize the local variations in each ROI. A logistic regression model was used to evaluate the predictive ability of this method. These results demonstrate the potential of TBT analysis to predict the incidence of knee OA using X-ray images. More specifically, the combination of TBT parameters with clinical covariates (age, sex, and BMI) significantly improved the KL prediction results from 0.57 AUC to 0.69 AUC.

In [Hladůvka et al., 2017], Hladůvka *et al.* investigated femoral textures as an indicator of OA risk and the potential of entropy as a computationally efficient alternative to texture descriptors. The proposed method is divided into three steps : ROI placement, feature extraction, and statistical analysis to evaluate feature combinations. First, a semi-automatic method was used to extract six ROIs from the femoral and tibial trabecular bones, as shown in Figure 2.8. The Hurst coefficient, H, Shannon entropy, E, and texture descriptors were computed for each ROI. Six ROIs using two descriptors yielded a set of 12 features for each subject. The final step involved studying the importance of each feature. This approach was evaluated by using 153 knee radiographs. The experimental results show that there is an indication of OA in the femur, in addition to well-known changes in the tibia. Furthermore, the best-selected feature set for OA prediction included the H coefficient of the medial tibia (T_0 and T_1) and entropy E of the medial femoral F_0 and lateral tibia (T_2 and T_3). The main limitation of this study was the moderately used sample size.

Riad *et al.* [Riad et al., 2018], present another approach based on complex wavelet decomposition for texture analysis of radiographic OA in knee X-ray images. This study used 688 X-ray images from the OAI database. Three square ROIs denoted as A, B, and C, each measuring 128 × 128 pixels, were

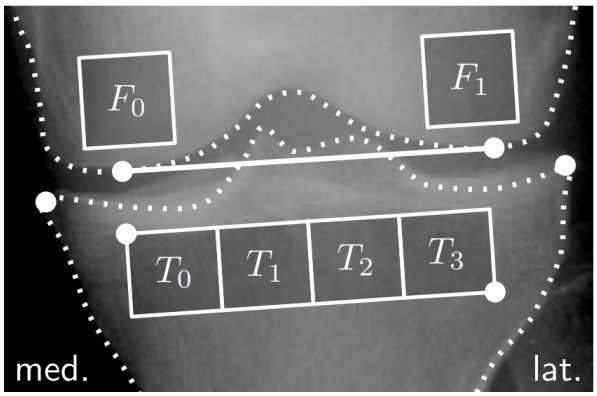


Figure 2.8: The six ROIs used in [Hladůvka et al., 2017]. Two ROIs (F_0 and F_1) represent the femoral part, and four (from T_0 to T_3) represent the tibial part.

selected from the medial, middle, and lateral parts of the tibia, respectively, as shown in Figure 2.9. The proposed method is divided into five main parts : First, an image pre-processing step based on a high-pass filter was introduced to retain crucial bone texture information in the selected ROIs. Second, an Undecimated Dual-Tree Complex Wavelet Transform was performed on each filtered ROI. Third, the relative phase was computed and modeled using two circular distributions : the Von Mises and the Wrapped Cauchy distributions. Fourth, the parameters of both models are estimated using the maximum likelihood estimator algorithm. Finally, the estimated parameters were used in the classification task to distinguish between normal knees (KL grade 0) and minimal OA (KL grade 2). The obtained classification results in terms of specificity, sensitivity, and accuracy were 85.47%, 75.29%, and 80.38%, respectively. These results show that the complex wavelet decomposition and statistics of the relative phase related to texture analysis are potentially powerful tools for OA detection compared to different feature extraction

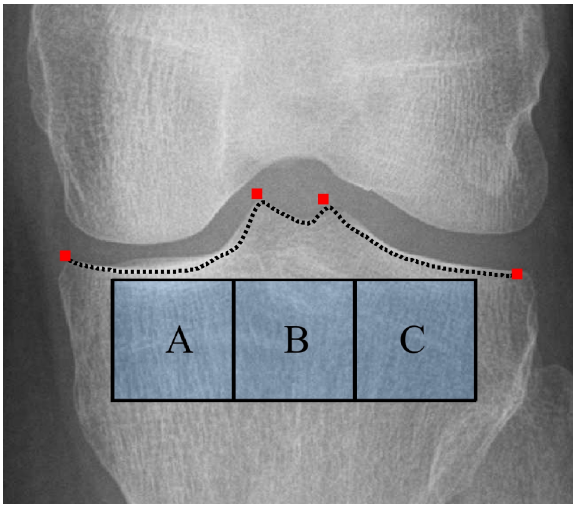


Figure 2.9: The ROIs used by Riad *et al.* [Riad et al., 2018]. The blue squares A, B, and C represent the medial, middle, and lateral of the tibia region, respectively.

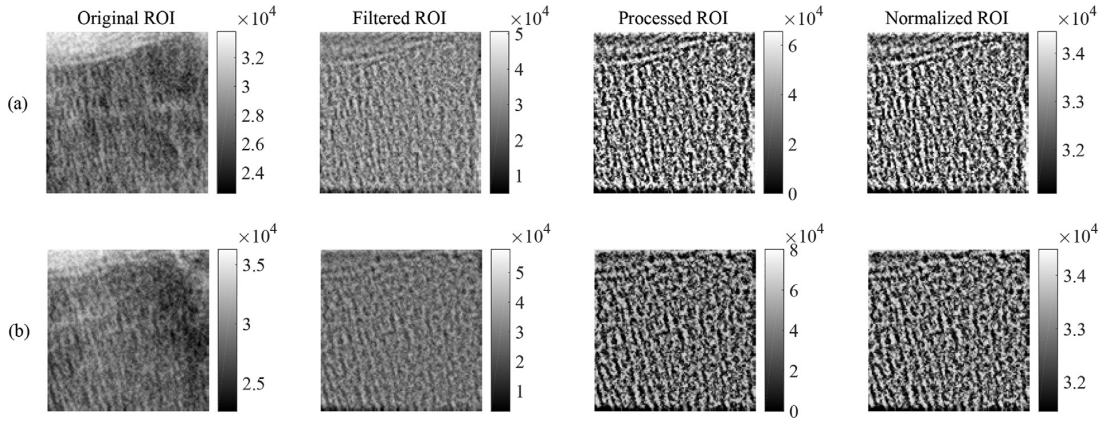


Figure 2.10: A representative ROI of the tibial trabecular bone and the different pre-processing applied in [Brahim et al., 2019]. ROI of (a) a control case subject and (b) an OA patient.

methods.

In [Brahim et al., 2019], Brahim *et al.* present a Computer-Aided Diagnosis (CAD) system for early knee OA detection using knee X-ray images and machine learning. Their proposed approach was applied to 1024 knee X-ray images from the OAI database, 514 knees from control subjects, and 514 knees from OA subjects. The proposed CAD system consists of several steps. First, three ROI on the medial, middle, and lateral sides of 128×128 pixels were extracted from each knee radiograph. These ROIs were then preprocessed in the Fourier domain using a circular filtering transform to stationarize X-ray images. Next, to retain the essential information of the data, a gray-level quantization is performed, followed by a new method of intensity normalization based on predictive modeling that uses Multivariate Linear Regression (MLR) to reduce the variability between OA and healthy subjects (see Figure 2.10). At the feature selection/extraction stage, an independent component analysis (ICA) approach was used to reduce the dimensionality. Finally, Naive Bayes and random forest classifiers are used in the classification task. The obtained results showed that normalization using MLR enabled not only to reduce the inter-subject variability but also to increase in the separation between the control and OA groups. In terms of performance, the proposed system achieved a good predictive classification rate for OA detection (82.98% for accuracy, 87.15% for sensitivity, and up to 80.65% for specificity).

2.3.1.B Detection approaches based on shape features

As shown in Table 2.1, a few studies have used only shape features to detect knee OA from X-ray images. In [Haverkamp et al., 2011], *Haverkamp et al.* investigated the role of joint shape in knee OA by identifying aspects of bone shape that differed in OA knee compared to control knees. This study used $1,218$ knee images of 609 women extracted from the Rotterdam study (RS-III-1) [Hofman et al., 2009]. To analyze knee shape, the authors used Statistical Shape Models (SSMs) consisting of 105 points to outline the contours of the femur, tibia, patella, and back of the medial condyle, as shown in Figure

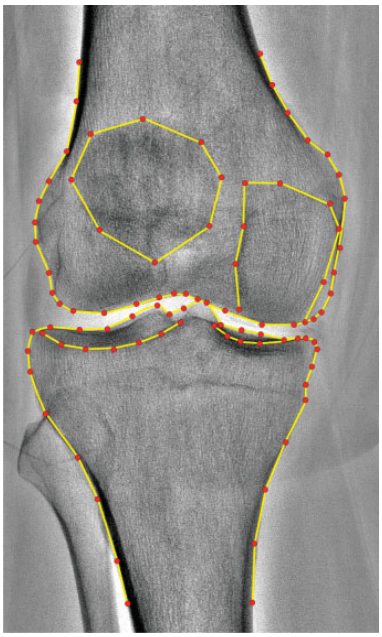


Figure 2.11: The 105 points considered by the statistical shape model [Haverkamp et al., 2011] to analyze the shape of the knee radiographs.

2.11. Then, a logistic Generalized Estimating Equation (GEE) regression model was used to analyze the association between various shape modes and the presence of OA (KL grade > 2). The results indicate that OA knees tend to have wider femoral and tibial bones relative to the shaft diameter, which increases with OA severity. These results show that the shape of the knee is involved in OA.

Minciullo *et al.* [Minciullo and Cootes, 2016] proposed an approach similar to that in [Haverkamp et al., 2011] for OA detection using lateral knee radiographs. They used a dataset including 300 lateral knee radiographs and 60 images per grade, extracted from the MOST study. For the binary classification task, the grades were split into non-OA (KL 0-1) and OA (KL 2-4). The proposed approach comprises two main stages : landmark point detection and OA classification. In the first step, a statistical shape model based on Principal Component Analysis (PCA) was applied to calculate the shape parameters (i.e., features).

2.3.1.C Detection approaches based on texture-shape features

The idea of using texture and shape features to characterize OA changes in X-ray images was first proposed by Shamir *et al.* [Shamir et al., 2008]. The proposed method first extracts a large set of shape, texture, and some statistical features from a set of 350 knee X-ray images taken from the Baltimore Longitudinal Study of Aging (BLSA) [Shock, 1984]. These features were computed from the raw pixels and several images transforms. In total, a features vector of 1470 was extracted. Then, a Fisher score selection technique was applied to reduce dimensionality and retain only the most informative image features.

Consequently, a feature space of 147 image features was maintained. Finally, the resulting feature vector was classified using the weighted nearest neighbor rule to predict the KL grade. The results indicate that more than 95% of moderate OA cases were accurately distinguished from normal cases. Furthermore, the classification accuracy for distinguishing between minimal OA and normal cases is approximately 80%. However, a classification accuracy of 54% was achieved for KL grade 1 (OA questionable) versus 0 (normal cases). This latter result is explained by the fact that these two degrees are so visually similar that even experienced human readers often struggle with the difference between them.

Another interesting study was proposed by [Thomson et al., 2015] by Thomson *et al.*. Their proposed approach combines both shape and texture features to better identify signs of OA disease from knee radiographs. This study used a set of 500 X-ray images distributed as follows : 100 KLG-0, 142 KLG-1, 82 KLG-2, 118 KLG-3, 43 KLG-4. First, each knee image was annotated with 74 landmark points (Figure 2.12). These annotations were then used to train a statistical shape model and a RFCLM object detection algorithm to extract the shape features. The second step was to select a ROI under the medial tibial plateau. This ROI was then analyzed to compute two texture measures : the Fractal Signature and Pixel Ratio Features. Experiments were performed to compare the classifiers using shape information, texture information, and both shapes and textures. The results show that combining shape and texture leads to a considerable improvement in the overall classification performance, with an AUC of 0.849 compared to 0.789 for shape and 0.754 for texture alone.

A more recent study [Minciullo et al., 2017] proposed combining features of lateral and PosteroAnterior (PA) view radiographs to better study the development of knee OA. This study used data from the MOST

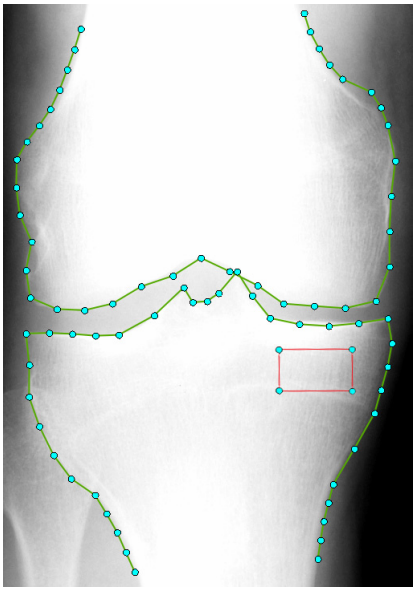


Figure 2.12: Example of the 74 landmark points and 4 points computed from them, outlining the ROI used for evaluating tibial texture used in [Thomson et al., 2015]

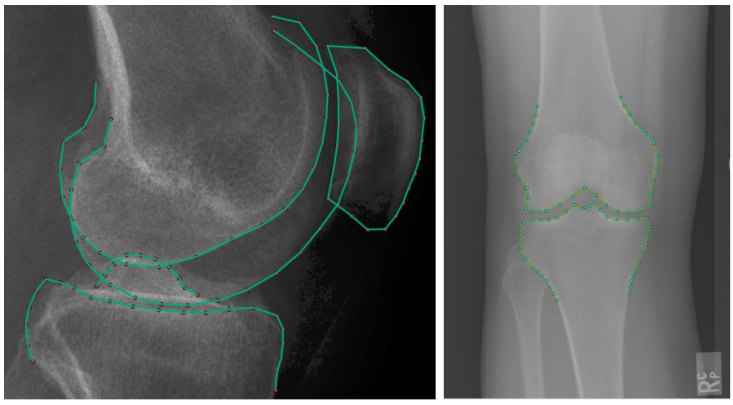


Figure 2.13: An example of the landmark points used to build the two shape models proposed in [Minciullo et al., 2017]

and OAI datasets and divided the KL grades into two groups : non-OA, KLG (0,1), and OA groups, KLG (2-4) to perform binary classification. The proposed method follows several steps for both lateral and PA radiographs. The first was to build a semi-automatic annotation model to detect the ROI containing the joint. ROIs detection was performed using an object detector based on Random Forests (RF) to initialize RFCLM automatically. The entire lateral knee annotation model is made of 102 landmark points, and the PA model is made of 74 points (see Figure 2.13). Once the annotation model was found, a set of shape, texture, and appearance parameters was extracted from each image and combined with the lateral and PA features to perform classification and prediction tasks. The experimental results show that the PA view has more discriminative features for classifying and predicting OA. Moreover, the combination of lateral and PA features led to a slight improvement in classification performance.

2.3.2 Deep learning-based approaches

Over the last few years, deep neural network architectures have gained remarkable attention from the computer vision research community and have achieved great success in various medical imaging applications, including detection, segmentation, and classification [Antony et al., 2017, Tiulpin et al., 2018, Chen et al., 2019, Norman et al., 2019, Nguyen et al., 2020]. Several studies using Deep Learning (DL) have recently been proposed for diagnosing knee OA using X-ray images.

In [Antony et al., 2017], Antony *et al.* employed the first deep Convolutional Neural Networks (CNN) to quantify radiographic knee Osteoarthritis severity using OAI and MOST datasets. A total of 8892 X-ray images were used from the OAI and distributed by KL grade as follows : KL-G0 - 3433, KL-G1 - 1589, KL-G2 - 2353, KL-G3 - 1222, and KL-G4 - 295. A total of 5840 knee X-ray images were selected from the MOST dataset and distributed as follows : KL-G0 - 2498, KL-G1 - 1018, KL-G2 - 923, KL-G3 - 971, and KL-G4 - 430. Figure 2.14 shows the proposed pipeline for quantifying the severity of knee OA. The first step of the proposed method is to detect the knee joint region (300 $\times$ 300 pixels). To this end,

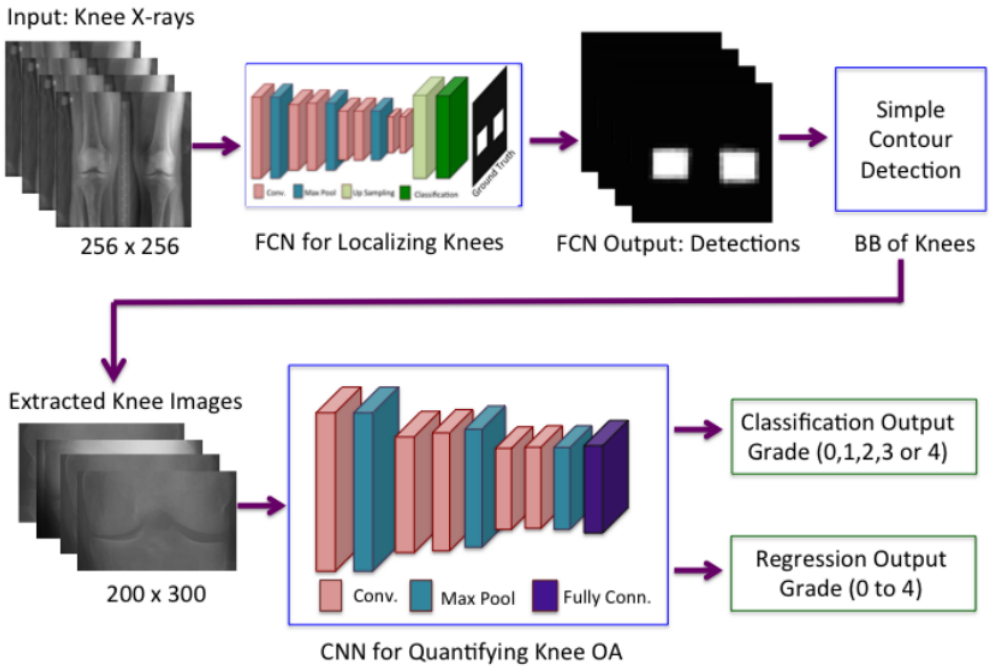


Figure 2.14: The proposed pipeline by Antony *et al.* [Antony et al., 2017] for quantifying knee OA severity

the authors trained a convolutional neural network to further improve the accuracy and precision of knee joint detection. Two approaches were used to classify the severity of knee OA : 1. Training the CNN for classification; 2. Jointly training CNN for classification and Regression. The experimental results showed that it is more appropriate to treat KL grade as a continuous variable and evaluate network performance using the mean square error. They also showed that the network trained jointly for classification and regression yields higher multiclass classification accuracy and lower mean-square error than the network trained only for classification. Moreover, the authors showed that distinguishing knee images of KL-G0, KL-G1, and KL-G2 is challenging because of the slight differences that do not appear between them.

Tiulpin *et al.* [Tiulpin et al., 2018] proposed an approach based on a deep Siamese CNN to diagnose knee OA from plain radiographs automatically. The authors used $18,376$ knee images from the MOST study and $8,917$ knee radiographs from the OAI dataset. Figure 2.15 shows the proposed classification pipeline for scoring knee OA severity. The method proposed in [Tiulpin et al., 2018] was used to localize the joint area on plain knee radiographs. Once the knee joint was localized, two patches corresponding to the medial and lateral knee sides were extracted to be analyzed using Siamese networks. They trained three models using different random seeds, and fused the predictions using a softmax layer to predict the resulting KL grade. In their study, the authors provided additional information using attention maps to show the areas of interest affecting the network decision.

Another interesting study was conducted in [Chen et al., 2019]. This approach fine-tuned several popular CNN models with a novel adjustable ordinal loss to classify the OA severity. Motivated by the

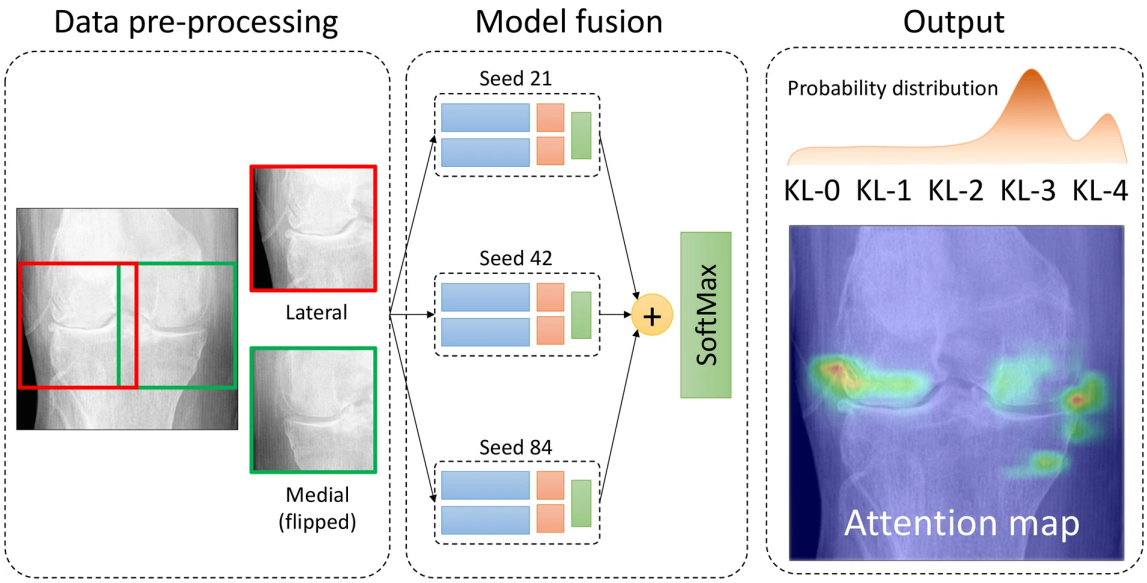


Figure 2.15: The classifcation pipeline proposed by Tiulpin *et al.*. [Tiulpin et al., 2018]

ordinal nature of the KL grading system, they assigned a higher penalty to incorrect classifications by imposing a larger distance between the predicted KL grades and real KL grades. Figure 2.16 illustrates the proposed pipeline. First, to detect the knee joint, the authors customized an existing CNN called YOLOv2. Second, to classify the detected knee joint images, they replaced the standard cross-entropy loss with an ordinal loss to fine-tune several popular CNN models, including the variants of ResNet, VGG, DenseNet, and InceptionV3. The baseline OAI dataset is used to train and evaluate the models. The experimental results showed that using ordinal loss improves the classification performance and reduces the mean absolute error between the predicted and true labels compared to the cross-entropy loss.

In [Nguyen et al., 2020], the authors proposed a new semi-supervised method (*Semixup*) for the automatic KL grading of knee OA using plain radiographs. Their method used a *mixup* technique to

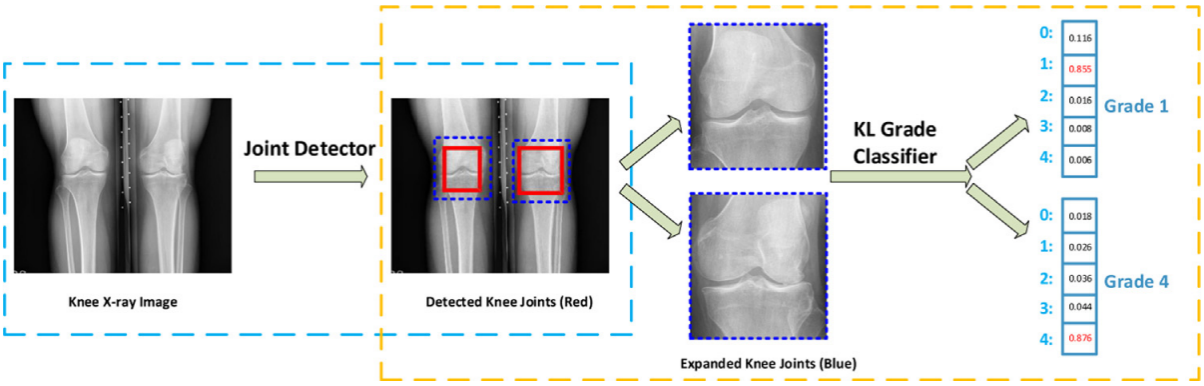


Figure 2.16: Illustration of the knee joint severity grading pipeline proposed in [Chen et al., 2019]. It includes knee joint detection and knee KL grade classification

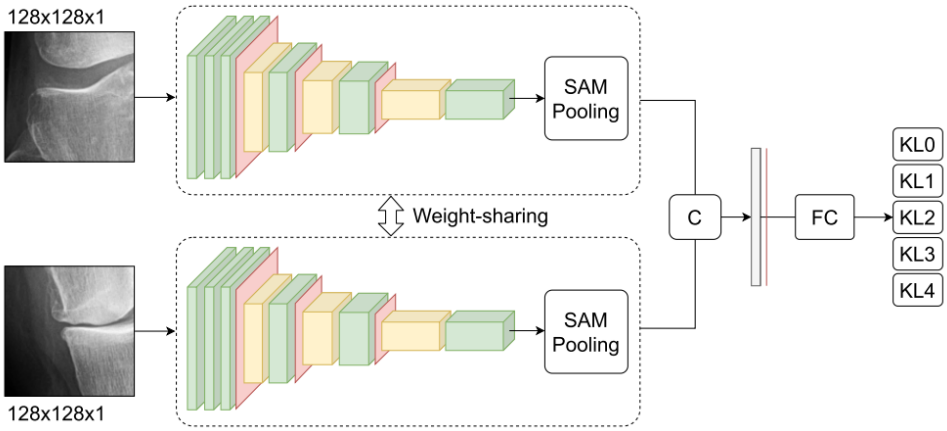


Figure 2.17: The Siamese model used by Nguyen *et al.* [Nguyen et al., 2020] to predicts KL grades corresponding to the inputs knee joint sides (lateral and medial).

generate out-of-manifold samples close to the data manifold. The architecture of the network (Figure 2.17) was based on the Siamese model previously developed by Tiulpin *et al.* [Tiulpin et al., 2018]. The main difference is the use of a Separable Adaptive Max-Pooling (SAM) and a fully connected layer instead of Global Average Pooling (GAP), which are concatenated and fed directly to the softmax layer. SAM pooling is based on applying pooling along one direction (horizontal or vertical) of the feature map. This pooling technique enables large feature maps to be processed efficiently and thus takes advantage of the accurate information required for the prediction of knee OA. The Siamese network was trained in a semi-supervised manner using an existing technique called *mixup* [Zhang et al., 2017]. For training and validation, the authors used OAI and MOST for testing. The results showed that the performance of the *Semixup* model improved significantly with the amount of training data and outperformed its corresponding supervised model when using at least 100 samples for each KL grade. In their study, the authors note performance saturation when the number of labeled training data exceeds $1,000$ images per KL grade.

Recently, Konwer *et al.* [Konwer et al., 2022] proposed a new paper discusses how clinical outcome or severity prediction from medical images has largely focused on learning representations from single-timepoint or snapshot scans. Motivated by the studies shown that disease progression can be better characterized by temporal imaging. The authores presented a deep learning approach that leverages temporal progression information to improve clinical outcome predictions from single-timepoint images. Figure 2.18 illustrate the proposed network. In their method, a self-attention based Temporal Convolutional Network (TCN) was used to learn a representation that is most reflective of the disease trajectory. Meanwhile, a Vision Transformer is pretrained in a self-supervised fashion to extract features from single-timepoint images. The key contribution is to design a recalibration module that employs maximum mean discrepancy loss (MMD) to align distributions of the above two contextual representations. The proposed

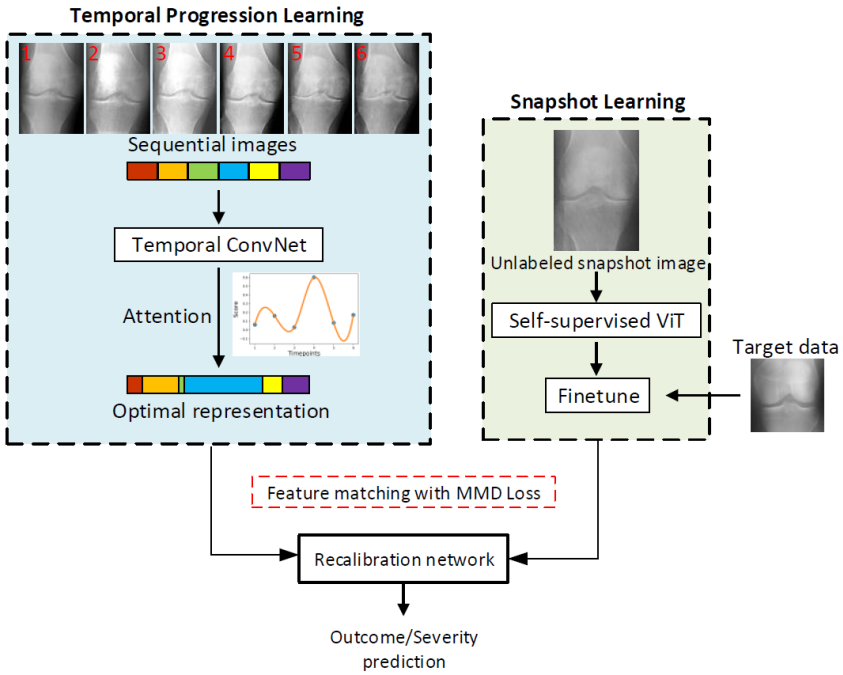


Figure 2.18: Overview of the framework proposed in [Konwer et al., 2022]. The temporal learning CNN module learns the optimal representation from sequential images. Self-supervised ViT extracts representations from snapshot images. The Recalibration network aligns the snapshot and temporal representations using MMD loss.

method was trained to predict clinical outcomes and severity grades from single-timepoint images. Experiments on osteoarthritis radiography datasets demonstrate that their approach outperforms other state-of-the-art techniques.

Table 2.2: SUMMARY OF CLASSIC CLASSIFICATION STUDIES OF KNEE OA BASED ON THE EXTRACTION OF **TEXTURE** FEATURES FROM RADIOGRAPHIC IMAGES

Author,Year	Images per grade	Feature enginering	Used Model	Results	Advantages	Desadvantages
[Woloszynski et al., 2012]	KL-G0 : 68 KL-G2 : 135	Roughness, degree of anisotropy, and direction of anisotropy	Logistic regression	AUC of 74% for medial ROI and of 68% for lateral ROI	The exracted features are invariant to a range of image magnification, exposure, noise, and blur	- Too small Sample size - The extracted features do not provide information at individual scales
[Hirvasniemi et al., 2014]	KL-G0 : 110, KL-G1 : 28 KL-G2 : 27, KL-G3 : 31 KL-G4 : 7	Entropy and homogeneity index of the LBP-based and Laplacian-based image	Linear mixed model	Most significant changes were seen in the medial subchondral bone plate and trabecular bone in proximal tibia and in medial femur	Analysis of all of the subchondral bone in the tibia, femur and tibial trabecular bone	The extracted texture parameters can be affected by the heel effect
[Janvier et al., 2017]	KL-G0 : 265 KL-G1 : 16 KL-G2 : 39 KL-G3 : 22 KL-G4 : 2	Fractal analysis	Logistic regression	AUC of 69%	A detailed statistical analysis was performed	This study has not completely investigated the piecewise fractal aspect of the trabecular bone texture and a fortiori cut-off scale
[Hladůvka et al., 2017]	KL-G$\leq$1 : 67 and KL-G$\geq$2 : 86	Hurst coefficient and Shannon entropy	Linear SVM	AUC of 85%	Use a mixture of femur, entropic and standard texture descriptors	Small sample size
[Riad et al., 2018]	KL-G0 : 344 KL-G2 : 344	Undecimated Dual Tree Complexes Wavelets Transform	SVM RBF	Accuracy (Acc) of 80.38% Sensitivity (Sen) of 85.47% Specificity (Spe) of 75.29%	The use of a new concept of the relative phases of complex coefficients	lack of statistical analysis
[Brahim et al., 2019]	KL-G0 : 514 KL-G2 : 514	Independent Component Analysis (ICA)	Random Forest	Acc of 82.98%, Sen of 87.15% and Spe 80.65%	The use of MLR normalization to reduce the intersubject variability	Not powerful enough to capture capture complex associations between raw data

Table 2.3: SUMMARY OF CLASSIC CLASSIFICATION STUDIES OF KNEE OSTEOARTHRITIS BASED ON THE EXTRACTION OF **SHAPE** FEATURES FROM RADIOGRAPHIC IMAGES

Author,Year	Images per grade	Feature enginering	Used Model	Results	Advantages	Desadvantages
[Haverkamp et al., 2011]	KL-G0 : 911 KL-G1 : 223 KL-G2 : 52 KL-G≥3 : 20	Statistical shape model based on PCA	Logistic regression	Knee shape is involved in OA	Well-detailed statistical analysis	-Study OA in women only - No classification test was performed
[Minciullo and Cootes, 2016]	KL-G≤1 : 120 and KL-G≥2 : 180	Statistical Shape Model (PCA)	Random Forests	AUC of 84.20% and 47.90% for binary and multi-class classification, respectivly	Study the lateral view of the knee X-ray	Low multiclass classification results

Table 2.4: SUMMARY OF CLASSIC CLASSIFICATION STUDIES OF KNEE OSTEOARTHRITIS BASED ON THE EXTRACTION OF **SHAPE-TEXTURE** FEATURES FROM RADIOGRAPHIC IMAGES

Author,Year	Images per grade	Feature enginering	Used Model	Results	Advantages	Desadvantages
[Shamir et al., 2008]	KL-G0 : 154 KL-G1 : 102 KL-G2 : 39 KL-G3 : 55	Set of image descriptors and image transforms (wavelet, Fourier, and Chebyshev)	Weighted KNN	KL-G0 vs. KL-G1 : Acc of 54% KL-G1 vs. KL-G2 : Acc of 60% KL-G0 vs. KL-G2 : Acc of 80% Multi-class include KL-G3 : Acc of 47%	Use a selection of texture and shape descriptors	- Poor resolution radiographs - Low multiclass classification results
[Thomson et al., 2015]	KL-G≤1 : 256 and KL-G≥2 : 244	Fractal Signature Statistical Shape Model	Random Forest	Texture : AUC of 75.40% Shape : AUC of 78.90% Combined : AUC of 84.50%	Fully automated method	- Considered KL-G1 as a Non-OA - Lack of evaluation experiences
[Minciullo et al., 2017]	KL-G≤1 : 6805 and KL-G≥2 : 4682	Combined Appearance Models [Cootes et al., 2004]	Random Forest	AUC of 85.30% for Knee Lateral view AUC of 90.40% for Knee PA view AUC of 90.50% for Knee Lateral+PA views	Combined shape and texture features of Lateral and PA view	- Considered KL-G1 as a Non-OA - Simple combination method

Table 2.5: DEEP LEARNING-BASED CLASSIFICATION STUDIES OF KNEE OA FROM RADIOGRAPHIC IMAGES

Author,Year	Images per grade	Used Model	Results	Advantages	Desadvantages
[Antony et al., 2017]	KL-G0 : 5931 KL-G1 : 2607 KL-G2 : 3276 KL-G3 : 2193 KL-G4 : 725	CNN	Acc of 60.30% and MSE of 89.80% Pr of 61% and Re of 63% F1 of 61%	CNN was jointly trained on classification and regression	Moderate classification results
[Tiulpin et al., 2018]	KL-G0 : 10954 KL-G1 : 4640 KL-G2 : 5430 KL-G3 : 4538 KL-G4 : 1731	Siamese CNN based on ResNet-34	Acc of 66.71% and AUC of 93%	Provide additional information such as the attention maps and the probability for specific KL grades	Poor classification accuracy for the most relevant OA cases KL-G1 (45%) and KL-G2(52%)
[Chen et al., 2019]	KL-G0 : 3195 KL-G1 : 1480 KL-G2 : 2235 KL-G3 : 1115 KL-G4 : 225	VGG-19	Acc of 69.63% and MAE of 35.80%	Propose a novel ordinal loss to consider the ordinal nature of the KL knee grading task	Very low classification accuracy for KL-G1 (17.56%)
[Norman et al., 2019]	KL-G0 : 16044 KL-G1 : 7514 KL-G2 : 9421 KL-G3 : 5124 KL-G4 : 1490	Ensemble of DenseNet	Acc of 70.98%	The use of a large amount of data	Consider KL-G1 as no-OA
[Nguyen et al., 2020]	KL-G0 : 17504 KL-G1 : 8204 KL-G2 : 10137 KL-G3 : 5787 KL-G4 : 1715	Siamese CNN	Balanced Accuracy of 71%	Achieve an accurate KL grading by using only small amounts of labeled data	Low classification accuracy for KL-G0 (69%) and KL-G1 (38%)

2.4 Summary

In this chapter, we begin by providing an overview of knee osteoarthritis, including its prevalence, impact, and current methods of diagnosis. We then introduce the OAI and MOST datasets, which were used in our experiments and are widely used in the field of knee osteoarthritis research. The focus of this chapter is on presenting various approaches for diagnosing early stages of knee osteoarthritis. We conduct a structural literature review, dividing the approaches into two main categories: hand-crafted feature-based methods and deep learning-based methods. Within each of these categories, we further divide the approaches into subcategories based on the type of descriptor used in the case of classical methods (texture, shape, shape-texture) and the type of model employed in the case of deep learning methods (AE or CNN). Finally, we present the results of our analysis in the form of four tables, which compare the performance, avantages and desavantages of classical methods based on texture (Table 2.2), shape (Table 2.3), shape-texture (Table 2.4), and deep learning methods (Table 2.5). The results show that deep learning methods generally outperform the classical methods, with the shape-texture and deep learning approaches exhibiting particularly strong performance compared to the texture and shape-only methods. This suggests that deep learning models using both shape and texture information may be particularly effective in accurately diagnosing early knee osteoarthritis.

3

Deep Learning Overview

Contents

3.1 Introduction

DEEP learning research has seen remarkable progress over the past decade in many areas, particularly in its applications in medical imaging. Deep learning is a subfield of machine learning methods based on artificial neural networks. Without human intervention, these neural networks learn to perform complex tasks directly from data. To understand the idea behind deep learning and the methods proposed in this thesis, it is necessary first to introduce the fundamental concepts of deep neural networks. This chapter presents a general overview of feedforward neural networks, deep representation learning, and learning techniques such as regularization. In particular, we introduce the AutoEncoder network, the backbone of the model presented in Chapter 4. This work was published in the journal IEEE-Transactions on Medical Images [Nasser et al., 2020]. In addition, we present advanced Convolutional Neural Networks related to the work presented in Chapter 5.

3.2 Feedforward neural networks

3.2.1 Perceptron network

Perceptron [McCulloch and Pitts, 1943] is a mathematical model of a biological neuron. The perceptron network is the most basic form of a feedforward neural network and is used for supervised learning of binary classifiers. As illustrated in Figure 3.1, the perceptron consists of four main parts : input values x, weights w, bias w_0, net sum, and an activation function f. For an input with I dimensions, the output value $\hat{y}$ is generated as follow :

$$\hat{y} = f(x, w) = f(\sum_{i=1}^{I} w_i x_i) \tag{3.1}$$

The most basic form of an activation function is a simple binary function with only two possible results :

$$\hat{y} = f(x, w) = \begin{cases} 1 & \text{if } \sum_{i=1}^{I} w_i x_i > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

A continuous approximation of this Heaviside step function is a nonlinear function called the sigmoid function :

$$\hat{y} = f(x, w) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{I} w_i x_i\right)} \tag{3.3}$$

Given a set of training data $D_{train} = \{(x_k, y_k) | k = 1, \ldots, N\}$, we start the training process of the perceptron network by initializing each weight to zero or a small random value of approximately zero.

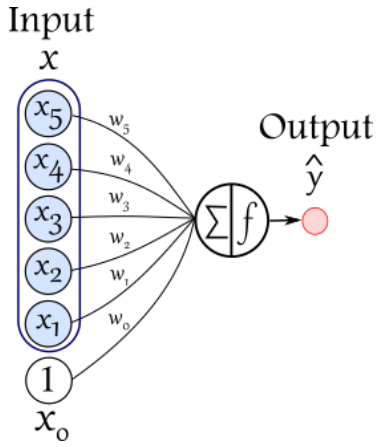


Figure 3.1: The perceptron network. The network use a set of weights w and an activation function f to map an input vector x to the output $\hat{y}$

Then, we take the inputs from a training set D_{train} and calculate the neuron's output $\hat{y}$. Next, we calculated the error, the difference between the neuron's output and the desired output in the training set example. Finally, depending on the direction of the error, we slightly adjust the weights according to the gradient descent rule :

$$w_i := w_i + \alpha(y - \hat{y})x \tag{3.4}$$

where, α is the learning rate. This process was repeated until the model converged. If the sigmoid function is used, the gradient descent update (Equation 3.4) is governed by the binary cross-entropy loss function :

$$J = -\frac{1}{N}\sum_{k=1}^{N} y_k \log(\hat{y}_k) + (1 - y_k)\log(1 - \hat{y}_k) \tag{3.5}$$

To deal with a multiclass dataset $D_{train} = \{(x_k, y_k)|k = 1, \ldots, N\}$ consisting of C distinct classes of data, a multiclass perceptron must be used. Where the output y_k is a one-hot coded vector of C units, in which $y_{ck} = 1$ if c is the class of training example k, and otherwise zero (see figure 3.2). The weight parameters were structured as a matrix $W \in R^{(I+1)\times C}$ rather than as a vector. For a training example x_k, the neuron output $\hat{y}_{ck}$ is calculated using the generalization of the sigmoid function for multiple classes as :

$$\hat{y}_c = f(x, W) = \frac{\exp\left(-\sum_{i=1}^{I} w_{ic}x_i\right))}{\sum_{\hat{c}=1}^{C} \exp\left(-\sum_{i=1}^{I} w_{i\hat{c}}x_i\right))} \tag{3.6}$$

The predicted class is then given by :

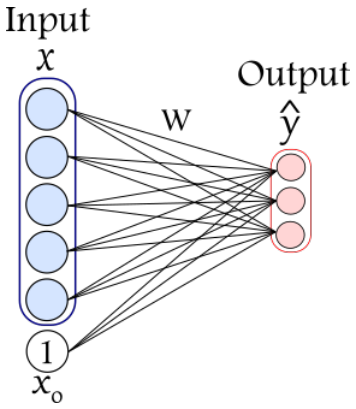


Figure 3.2: The multiclass perceptron network.

$$\hat{y}_k = \arg\max_{c=1,..,C} \hat{y}_{ck} \tag{3.7}$$

The objective function to be optimized for training the multiclass perceptron network is the general cross-entropy loss function :

$$J = -\frac{1}{N} \sum_{c=1}^{C} \sum_{k=1}^{N} y_{ck} \log(\hat{y}_{ck}) \tag{3.8}$$

Each weight parameter $w_{i,c}$ between the input unit i and the output unit c, is updated according to the following equation :

$$w_{ic} := w_{ic} + \alpha \frac{\partial J}{\partial w_{ic}} \tag{3.9}$$

3.2.2 Multi-Layer Neural Networks and Backpropagation

A multi-layer neural network or Multi-Layer Perceptron (MLP) is an artificial neural network that can learn a nonlinear function approximator for classification or regression by training on a dataset. A MLP consists of three types of layers : an input layer to receive the signal, an output layer that makes a decision or predicts the input, and a hidden layer between the input and output that represents the computation nodes of the MLP, as shown in Figure 3.3.

MLP training involves adjusting the model's parameters, or the weights and biases, to minimize errors. Backpropagation is used to make the weight and bias adjustments relative to the error, and the error itself can be measured by different loss functions J, such as cross-entropy loss or mean squared loss.

The MLP is a feedforward network. It mainly involves two motions : a constant back-and-forth motion. In the forward pass, the signal flow moves from the input layer through the hidden layers to the output layer, and the output layer decision is measured against the ground truth labels. In the backward

pass, using backpropagation and the chain rule of calculus, partial derivatives of the loss function with respect to the various weights and biases are backpropagated through the MLP. This act of differentiation gives a gradient along which the parameters may be adjusted as they move the MLP one step closer to the error minimum.

Consider a three-layer network (Figure 3.3) with the input layer $x \in R^I$ and hidden layer $h \in R^J$ linked by weights W_1, and the hidden layer linked to the output layer $haty \in R^C$ via weights W_2. The weights of the network are updated as follows :

$$\begin{aligned} w_{2,jc} &= w_{2,jc} + \alpha \frac{\partial J}{\partial w_{2,jc}}, \\ w_{1,ij} &= w_{1,ij} + \alpha \frac{\partial J}{\partial h_j} \frac{\partial h_j}{\partial w_{1,ij}} \end{aligned} \tag{3.10}$$

This update can be achieved with any gradient-based optimization algorithm such as stochastic gradient descent [Ruder, 2016]. The network continues to play the game of tennis until the error can no lower. This state is known as the convergence state.

A MLP with more than one hidden layer is called a deep neural network. Deeper neural networks are better at processing data. However, deeper layers can lead to vanishing gradient problems, and optimization becomes more difficult. Section 3.3 discusses particular methods to circumvent this issue.

3.2.3 Regularisation Techniques

A central problem in machine learning is how to create an algorithm that will perform well not only on the training data but also on new inputs. This problem is referred to as overfitting. Many strategies used in machine learning have been explicitly designed to reduce test errors, possibly at the expense of increased training errors. These strategies are collectively known as regularization.

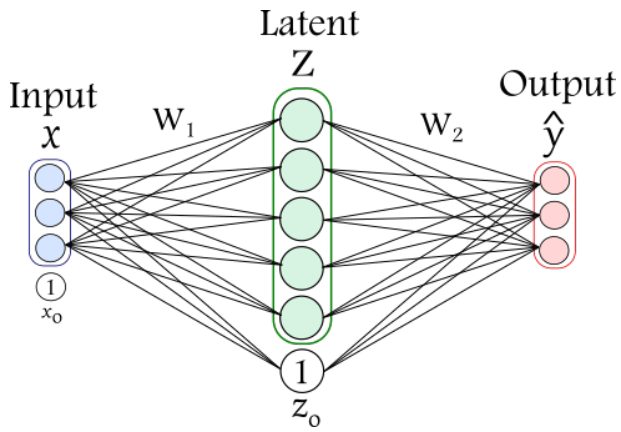


Figure 3.3: Multi-layer perceptron. Schematic representation of a MLP with single hidden layer.

Deep neural networks are complex learning models exposed to overfitting, owing to their flexible nature of memorizing individual training set patterns instead of taking a generalized approach towards unrecognizable data. Therefore, neural network regularization is essential. It helps to keep the learning model easy to understand, allowing the neural network to generalize data that cannot be recognized. Many forms of regularization are available to deep learning practitioners. The development of more effective regularization strategies has represented a significant research effort.

Regularization can be defined as any modification made to a learning algorithm to reduce its generalization error but not its training error. There are many regularization strategies. Some place extra constraints on the machine learning model, such as adding restrictions on the parameter values. Others add extra terms to the objective function that can correspond to a soft constraint on the parameter values. If chosen carefully, these extra constraints and penalties can lead to improved performance in the test set.

In a general learning algorithm, the dataset is divided into training and testing sets. After each epoch of the algorithm, the parameters were updated accordingly after understanding the dataset. Finally, the trained model is applied to the test set. The training set error was generally less than the test set error. This is owing to overfitting, in which the algorithm saves the training data and produces excellent results on the training set. Thus, the model is highly exclusive to the training set and fails to produce accurate results for other datasets, including the test set. Regularization techniques are used in such situations to reduce overfitting and increase the model's performance on any general dataset.

3.2.3.A L1 and L2 Regularization

L1 and L2 use the weight penalty regularization technique, which is commonly used to train models. It is assumed that models with larger weights are more complex than those with smaller weights. The role of the penalties is to ensure that the weights are either zero or very small. The weight Penalty, also known as Weight Decay, signifies the decay of weights to a smaller unit or zero.

During L1 and L2 regularization, the J loss function of the neural network is extended by the so-called regularization term Ω, which limits the model's capacity. We denote regularized objective function by $\tilde{J}$:

$$\tilde{J} = J + \lambda\Omega \tag{3.11}$$

where λ is a hyperparameter that weights the relative contribution of the regularization penalty term, Ω, relative to the standard objective function J. Setting λ to 0 results in no regularization. Larger values of λ correspond to greater regularization. The regularization term, Ω, differs between L1 and L2.

In L2, Ω is defined as :

$$\Omega = \frac{1}{2}\|W\|_2^2 \tag{3.12}$$

L2 regularization will significantly impact the weight vector's directions, which do not "contribute" much to the loss function. However, it had a relatively small effect on the directions that contributed to the loss function. As a result, we reduce our model's variance, making it easier to generalize to unseen data.

In the case of L1 regularization, we use another regularization term, Ω. This term is the sum of the absolute values of the weight parameters of the weight matrix :

$$\Omega = \frac{1}{2}\|W\|_1 \tag{3.13}$$

L1 regularization introduces sparsity in the weights by forcing more weights to be zero instead of reducing the average magnitude of all weights (as the L2 regularizer does). In other words, L1 suggests that certain features should be discarded during the training process.

3.2.3.B Data Augmentation

The best way to generalize a machine learning model is to train it using more data. Unfortunately, in practice, the amount of available data is limited. One way to solve this problem is to use data augmentation approaches to generate new data for the training set from the available ones.

Data augmentation can address the requirements, diversity, and amount of the training data. Additionally, it can be used easily to address the class imbalance problem in classification tasks. Images are high-dimensional and include an enormous range of variations, many of which can be easily simulated. Operations such as translating the training images can often greatly improve generalization, even if the model has already been designed to be partially translation invariant by using the convolution and pooling operations described in Section A.1. Many other operations, such as rotating or scaling the image, have also been used and proven quite effective. However, care must be taken not to apply transformations that would change the correct class. For example, optical character recognition tasks require recognizing the difference between *b* and *d* and the difference between *6* and *9*; thus, horizontal flips and 180-degree rotations are not appropriate ways to augment datasets for these tasks.

Injecting noise into the input of a neural network can also be considered a form of data augmentation. Input noise injection is part of unsupervised learning algorithms, such as the denoising autoencoder (see Section 3.3.2.B). Noise injection can also be applied to hidden units. This can be seen as an increase in the datasets at multiple levels of abstraction. Dropout is another powerful regularization strategy, which will be described in Section 3.2.3.C, and can be seen as a process of constructing new inputs by multiplying by noise.

When comparing the machine learning benchmark results, it is important to consider the effect of increasing the dataset. Often, hand-designed dataset augmentation schemes can significantly reduce the generalization error of machine learning techniques. To compare the performance of a machine learning algorithm A and another machine learning algorithm B, both algorithms should be evaluated by hand using the same designed dataset augmentation schemes. Suppose Algorithm A performs poorly without increasing the dataset, and Algorithm B performs well when combined with many synthetic input transformations. In such a case, the synthetic transformations likely caused performance improvement rather than the use of machine learning Algorithm B.

3.2.3.C Dropout

Dropout is another powerful regularization method that approximates the training of many neural networks with different architectures in parallel. Dropout is used to regularize a large family of deep neural networks.

During the training, some units on the dropout layer are randomly turned off (dropped out) with probability p, meaning fewer neurons work in the forward process. Thus, the overall structure of the neural network was simplified. In the test phase, however, we keep all units, so the values will be much higher than expected. Therefore, we must scale them down by p. The probability p of selecting the number of nodes to be dropped out is a hyperparameter of the dropout function. By using the dropout technique, the same layer alters its connectivity and searches for alternative paths to convey the information in the next layer. In other words, by applying the dropout, the model can no longer rely on specific neurons (as they could be muted in the process), and all other neurons would need to learn in the training phase. Therefore, each update of the dropout layer during training was performed with a different "view" of the configured layer. Conceptually, it approximates the training of many neural networks with different architectures in parallel, making the trained model more robust.

3.2.3.D Early stopping

When training a deep network with a significant learning capacity, the model attempts to steadily decrease the loss function of the training data. However, there is a point during training when the model stops generalizing and starts learning the statistical noise in the training dataset (overfitting the task). The challenge is to train the network sufficiently long to create good input-to-output mapping but not to train it so long that it overfits the training data. One approach to overcome this problem is to use early stopping regularization. This can be achieved by dividing the dataset into training, validation, and test sets. By using early stopping, the algorithm is trained on the training set, and the point at which training is stopped is determined from the validation set. This means that during training, we store a copy of the model parameters whenever the error in the validation set is improved. When the training

algorithm terminates, we return these parameters, which give the least validation set error rather than the latest parameters. The algorithm terminates when no parameters have improved over the best-recorded validation error for a pre-specified number of iterations.

Early stopping is one of the most commonly used forms of regularization in deep learning. Its popularity is owing to its effectiveness and simplicity. Early stopping can be considered implicit regularization in that it requires almost no change in the underlying training procedure, the objective function, or the set of allowable parameter values. Early stopping may be used alone or in conjunction with other regularization strategies. Early stopping is also useful because it reduces the computational cost of the training procedure.

Early stopping requires a validation set, which implies that some data are not fed into the model. To best exploit these extra data, additional training can be performed after the initial training has been completed. All the training data were included in the second training step. Two basic strategies were used in the second training procedure. One strategy is to initialize the model again and retrain all data. In this second training pass, we trained the network for the same number of steps as the early stopping procedure determined to be optimal in the first pass. However, attention must be paid to some subtleties associated with this strategy. For example, in the second round of training, each pass through the new dataset required more parameter updates because the training set was larger. Thus, there is no good way to determine whether retraining with the same number of parameter updates or the same number of steps will result in a better model.

Another strategy for using all the data is to retain the parameters obtained from the first round of training and continue training but now use all the data. At this stage, we no longer have a guide for when to stop. Instead, we can monitor the average loss function on the validation set and continue training until it falls below the value of the training set objective, at which the early stopping procedure is halted. This strategy avoids the high cost of retraining the model from scratch but is not as well-behaved. For example, the objective of the validation set may not reach the target value; therefore, so this strategy is not guaranteed to terminate.

3.3 Learning deep representations

In this section, we will introduce in detail the most basic deep neural networks and their learning process, namely 1) auto-encoder networks (3.3.1), 2) auto-encoder networks (3.3.1), regularized encoder (3.3.2) and ii) Convolutional Neural Networks (3.3.3). We will also present recent advances in the CNN network, including Discriminative CNNs (3.3.3.B) and Texture CNNs (3.3.3.C). These deep networks focus on training high-level representations that describe the input examples and capture the complex structure of the raw input data through a sequence of layers. This hierarchical structure of deep networks is inspired by the functioning of our brains. Thus, this deep neural network tries to simulate the behavior of the

human brain.

3.3.1 Autoencoder networks

Auto-Encoder (AE) neural network is an unsupervised learning algorithm that aims at reconstructing the output from the input [Bourlard and Kamp, 1988]. It consists of an encoder and a decoder part, as shown in Fig. 3.4. The former is employed to encode the input data into a latent-space representation (features), while the latter is used to reconstruct the encoded data.

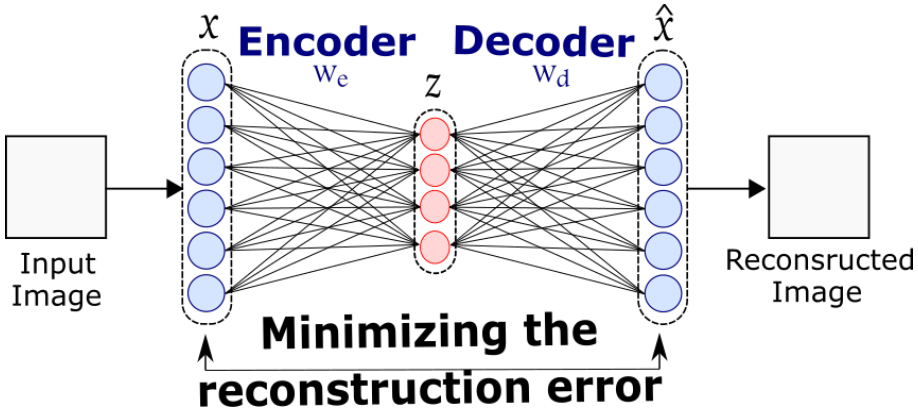


Figure 3.4: Illustration of the traditional AE network. The network learns useful properties of its input data by using the input reconstruction errors to update its parameters.

To encode an input vector $x \in \mathbb{R}^I$, the encoder maps this input linearly with a set of encoding weights $W_e \in \mathbb{R}^{K \times I}$ with K units. A bias vector $b_e \in \mathbb{R}^K$ was then added, and a nonlinear encoding activation function σ was applied to produce the outputs of the latent representation $z = f_e(x) = \sigma(W_e \cdot x + b_e) \in \mathbb{R}^K$. To obtain the reconstructed input $\hat{x} \in \mathbb{R}^I$, the decoder maps the encoded representation z with another set of decoding weights $W_d \in \mathbb{R}^{I \times K}$ as $\hat{x} = f_d(z) = \sigma(W_d \cdot z + b_d) \in \mathbb{R}^I$, where f_e and f_d are the encoder and decoder functions, respectively.

In its original form, the AE learns features by minimizing the reconstruction error, L, between input x and its decoded version $\hat{x}$. The cost function commonly used for optimization during the learning process is the Mean Square Error (MSE) [Ng et al., 2011, Bourlard and Kamp, 1988], J_{AE} defined as :

$$J_{AE}(\theta) = L(x, \hat{x}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} (x_{i,n} - \hat{x}_{i,n})^2 \tag{3.14}$$

where N is the number of training samples and θ represents the vector of all parameters of the network $\{W_e^l, W_d^l, b_e^l$ and b_d^l for $l = 1, ..., L\}$.

Traditionally, autoencoders have been used for dimensionality reduction and feature learning. Dimensionality reduction is performed using an **undercomplete AE**, in which the dimension of the latent layer is less than that of the input layer (i.e., $K < I$). Learning an undercomplete representation forces

the autoencoder to capture the most salient features of the training data. Learning an AE for feature encoding often requires a larger latent layer (i.e., $K > I$). However, this network, called an **overcomplete AE**, can potentially learn trivial solutions, such as the identity function. Moreover, an AE with nonlinear encoder functions, f_e, and nonlinear decoder functions, f_d, can learn more powerful features. Unfortunately, this autoencoder fails to learn anything useful if the encoder and decoder are given an excessive capacity.

3.3.2 Regularized Autoencoders

Undercomplete autoencoders can learn the most salient features of the data distribution in many cases. However, we have observed that these autoencoders do not learn anything useful if the encoder and decoder are given too much capacity. A similar problem occurs in the overcomplete case; even if we use a linear encoder and decoder, the overcomplete AE can learn to copy the input to the output without learning anything useful about the data distribution. Thus, an ideal autoencoder model balances the following :

- Sensitive to inputs sufficient to accurately build a reconstruction.
- Insensitive enough to inputs so that the model does not simply memorize or overfit the training data.

Regularized autoencoders provide this ability to do so. Rather than limiting the model capacity by using a linear encoder and decoder and keeping the code size small, regularized autoencoders use a loss function that encourages the model to possess properties other than the ability to copy its input to its output. These other properties include the sparsity of their presentation, the smallness of the derivative of the representation, and robustness to noise or missing inputs. A regularized autoencoder can be nonlinear and overcomplete but still learn something valuable about the data distribution, even if the model capacity is sufficiently large to learn a trivial identity function.

3.3.2.A Sparse Autoencoders

A Sparse AutoEncoder (SAE) is a popular regularized autoencoder that encourages sparsity on the latent representation by adding a regularizer term Ω to the reconstruction cost function :

$$\tilde{J} = J_{AE} + \lambda \Omega_{sparsity} \tag{3.15}$$

where λ denotes the coefficient of the sparsity regularization term. The regularizer term $\Omega_{sparsity}$ is a function of the average output activation value of a neuron. The average output activation measure of neuron i is defined as :

$$\hat{\rho}_i = \frac{1}{N}\sum_{n=1}^{N} z_i(x_n) = \frac{1}{N}\sum_{j=1}^{N} \sigma(w_e^{(i)} \cdot x_n + b_e^{(i)}) \tag{3.16}$$

where N denotes the total number of training samples. x_n is the n^{th} training example, $w_e^{(i)}$ is the i^{th} row of the encoding weight matrix W_e, and $b_e^{(i)}$ is the i^{th} element of the encoding bias vector b_e. Adding a term to the cost function that constrains the values of $\hat{\rho}$ to be low encourages a few nodes in the hidden layer to activate and makes the rest zero. Therefore, even if we have many hidden units (as in the overcomplete AE), it will only fire some hidden units and learn the useful structures present in the data.

We have observed that sparsity can be encouraged by adding a regularization term, $\Omega_{sparsity}$. This term should take a large value when the average activation value, $\hat{\rho}_i$, of neuron i and its desired value, ρ, are not close. One such sparsity regularization term is the Kullback-Leibler Divergence (KLD) [Ng et al., 2011].

$$\Omega_{sparsity} = \sum_{i=1}^{K} KL(\rho \parallel \hat{\rho}_i) = \sum_{i=1}^{K} \rho \log(\frac{\rho}{\hat{\rho}_i}) + \log(\frac{1-\rho}{1-\hat{\rho}_i}) \tag{3.17}$$

The KLD measures the difference between the two distributions. It takes a value close to zero when ρ and $\hat{\rho}_i$ are close to each other and become larger otherwise. Minimizing the cost function forces this term to be small; hence, ρ and $\hat{\rho}_i$ to be close to each other, resulting in a sparse representation Z.

3.3.2.B Denoising Autoencoders

Rather than constraining the representation by adding penalty Ω to the cost function, we can obtain an autoencoder that learns something useful by changing the reconstruction error term of the cost function.

We have seen that the reconstruction criterion alone cannot guarantee the extraction of valuable features as it can lead to learning an identity function "simply copy the input". The denoising AutoEncoder (DAE) avoids this problem by first corrupting the initial input x into $\tilde{x}$ and training it to reconstruct a clean input from this corrupted version (see Figure 3.5). Note that denoising autoencoders still minimize the same reconstruction loss between clean x and its reconstruction from $\hat{x}$. The difference is that $\hat{x}$ is obtained from corrupted input $\tilde{x}$ rather than x. Thus, it forces the DAE to undo this corruption by extracting useful features for denoising rather than simply copying their input.

3.3.2.C Contractive Autoencoder

The objective of the Contractive AutoEncoder (CAE) is to have a robust learned representation that is less sensitive to small variations in the data. This is accomplished by introducing an explicit regularizer $\Omega(z)$ on the code $z = f_e(x)$, encouraging the derivatives of f_e to be small as possible :

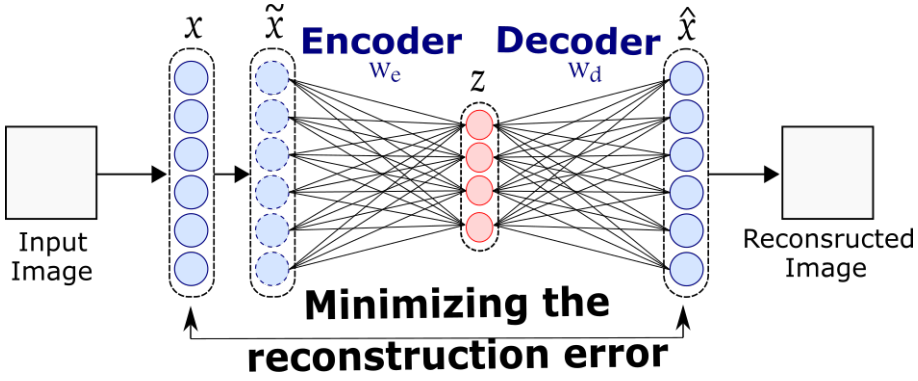


Figure 3.5: Illustration of the denoising autoencoder architecture. An input sample x is corrupted to $\tilde{x}$. The autoencoder then maps it to z (via encoder) and attempts to reconstruct x via decoder, producing reconstruction $\hat{x}$. Reconstruction error is measured by loss $L(x, \hat{x})$.

$$\Omega(z) = \lambda \left\| \frac{\partial f_e(x)}{\partial x} \right\|_F^2 \tag{3.18}$$

This penalty $\Omega(z)$ must conform to the squared Frobenius norm of the Jacobian matrix for the encoder function f_e with respect to the input x.

The CAE was trained to resist perturbation of its input. Thus, there is a connection between the denoising autoencoder and the contractive autoencoder. With a slight Gaussian noise in the input, the denoising objective function is equivalent to a contractive penalty on the reconstruction function that maps x to $\hat{x} = f_d(f_e(x))$.

The CAE penalty term generates a mapping that strongly contracts data. In other words, the CAE is trained to resist perturbations of its input, and it contracts the input neighborhood to a smaller output neighborhood; hence, it is called a contractive autoencoder.

3.3.2.D Deep Autoencoder

A deep autoencoder is an autoencoder with multiple hidden layers representing the encoder part of the network and its symmetric layers making the decoder part. The choice of the number of hidden layers depends on the input data's complexity. Using more than one hidden layer with nonlinear encoding and decoding functions allows the deep autoencoder network to capture a more complex relationship between the input data and thus extract the features most effectively. Figure 3.6 represents the structure of a typical deep autoencoder.

3.3.2.E Stacked Autoencoders

In some cases, the input data has a complex structure that is difficult to capture using a single autoencoder. To address this, stacked autoencoders can be used, which are multiple autoencoders arranged one

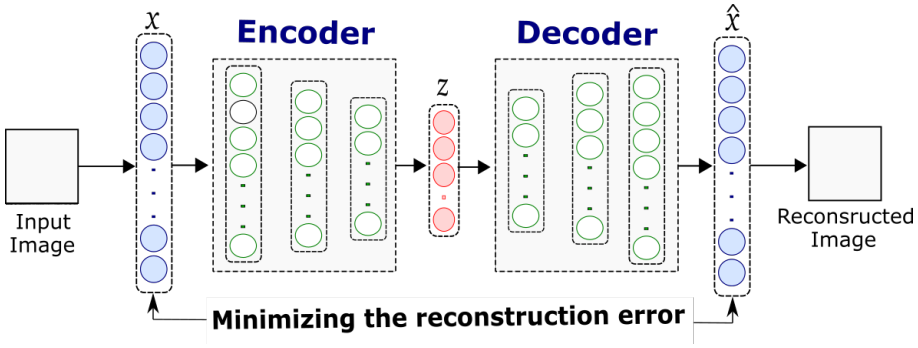


Figure 3.6: Illustration of a deep autoencoder's structure

on top of the other as shown in Figure 3.7. Stacking multiple autoencoders allows them to learn more complex codings and can be effective for certain tasks.

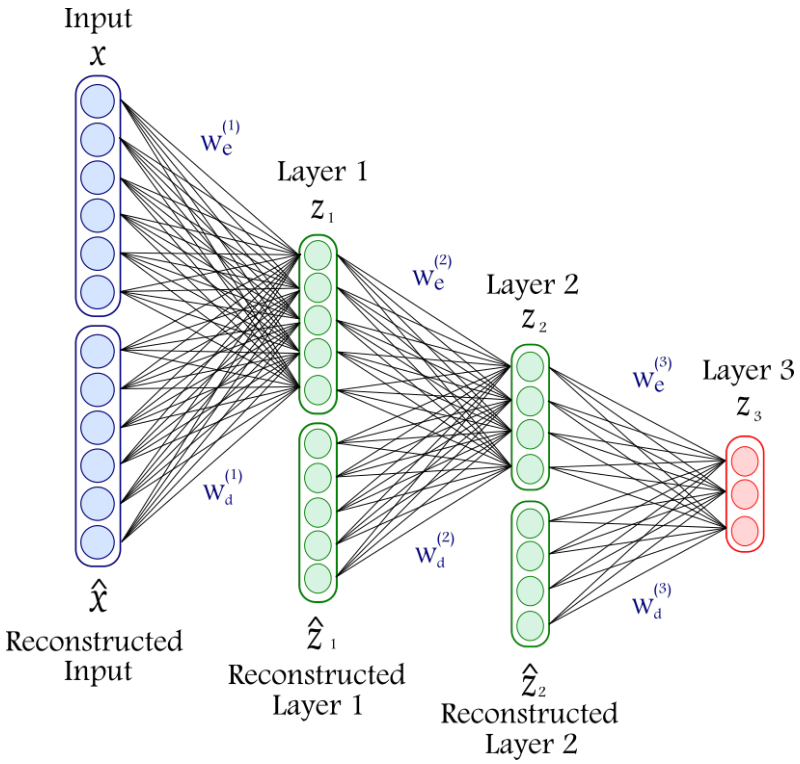


Figure 3.7: Stacked Autoencoder network. The network learns a sequence of autoencoders that attempt to reconstruct the previous layer via a second set of weights.

As illustrated in Figure 3.7, the stacked autoencoder consists of multiple autoencoders arranged one on top of the other, with the output of the hidden layer of each autoencoder connected to the input of the next. Each autoencoder attempts to minimize the reconstruction errors of the previous layer. The stacked autoencoder generally includes three steps :

- Train the first autoencoder using the input data, and obtain the learned feature vector.

- The learned feature from the trained autoencoder is used as an input for the next autoencoder, and which continues until training is completed.
- Once all the hidden layers are trained, backpropagation is used to minimize the cost function and update the weights with labeled training set to achieve fine-tuning.

3.3.3 Convolutional Neural Networks

3.3.3.A Overview and Brief history

In general, fully-connected deep architectures, such as Stacked and Deep AE shown above, perform extremely well on modeling input data. However, due to the fully-connected structure between layers, it is difficult to scale up to handle high-dimensional inputs, such as images. For example, an image of 224×224 pixels would have an input dimension of $50,176$ when flattened, which would have at least $50,176$ weights per neuron only in the input layer. This amount of weight rapidly becomes unmanageable for large images or when we want to add multiple hidden layers with varying nodes per layer.

A Convolutional Neural Network (CNN) is a specialized type of supervised feedforward neural network designed for computer vision tasks, such as analyzing 2D images, 3D volume data, and videos. CNNs were first introduced in 1989 by Yann LeCun [LeCun et al., 1989]. LeCun had built on the work of Kunihiko Fukushima [Fukushima and Miyake, 1982], who invented the Neocognitron, an elementary hierarchical multilayer neural network dedicated to image recognition.

In the CNNs architecture, not all nodes in layer l are connected to every node in the next layer, $l+1$. This means that the neurons of layers $l+1$ will only be connected to a small region of layer l. This principle, called local connectivity, is one of the critical concepts of CNNs, allowing the networks to reduce the trainable parameters significantly.

The name "convolutional neural network" indicates that the network employs a mathematical operation called convolution. This convolutional operator naturally shares the model's parameters by coding the image locally across various spatial locations in the image. This helps scale the model for large images while considering the spatial correlations in the image.

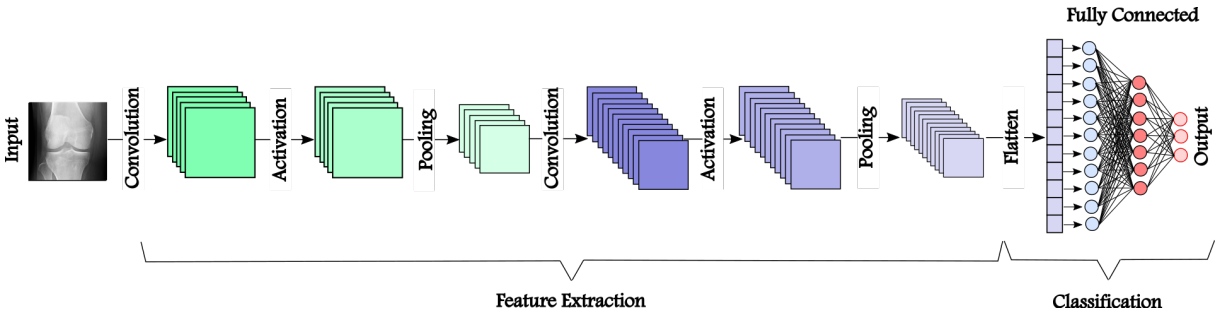


Figure 3.8: Basic architecture of a CNN

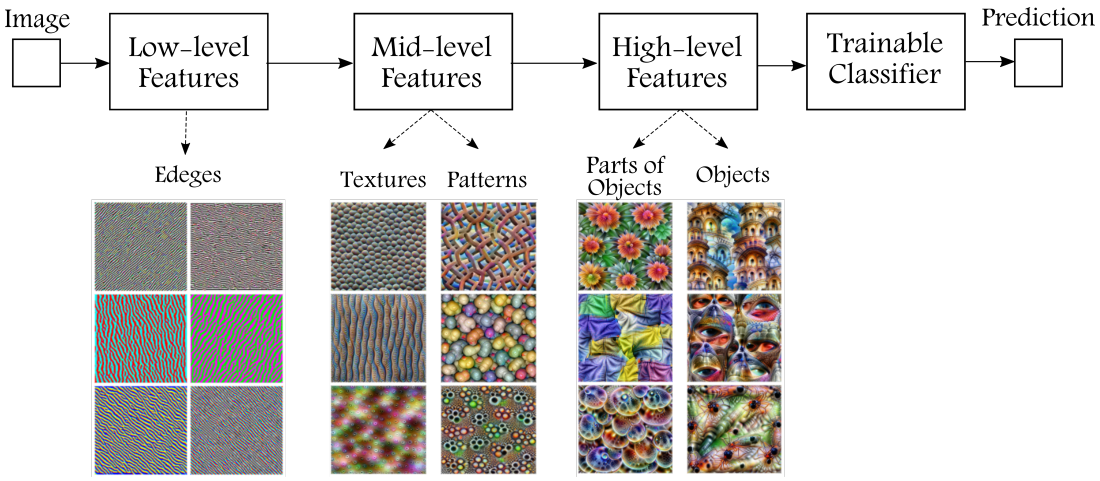


Figure 3.9: This figure illustrates the concept of abstraction levels in a convolutional neural network (CNN) architecture. The first layers of the CNN learn low-level features, such as edges. The later convolutional layers learn mid-level features, such as complex textures and patterns. The final layers learn high-level features, such as objects or parts of objects. The classifier, consisting of fully connected layers, uses the activations from the high-level features to predict the individual classes. This illustration is based on Olah et al. [Olah et al., 2017].

CNNs produce a series of transformations of the input image into latent representations until the final transformation produces the prediction. A basic CNN can be summarised by a succession of convolution, pooling, non-linear, and fully connected layers as shown in Figure 3.8. In those convolutional layers, the network learns new and increasingly complex features in its layers. Then the transformed image information goes through the fully connected layers and turns into a classification or prediction as illustrated in Figure 3.9.

In the following sections, we describe some advanced CNN models proposed in the literature related to our work presented in chapter 5.

3.3.3.B Discriminative CNNs

Several studies have demonstrated that when traditional convolutional neural networks (CNNs) are used with the softmax cross-entropy loss in case of data with high inter-class similarities or high intra-class variations. As a result, the features learned from the same class may be scattered, and those learned from different classes may overlap [Wen et al., 2016, Cai et al., 2018, Cheng et al., 2018].

Some discriminative CNN methods suggest using a new loss instead of the standard softmax loss : **Triplet Loss :** The triplet loss was first introduced in a deep convolutional network called FaceNet [Schroff et al., 2015]. It aims to ensure that an image referred to as the anchor is as close as possible to all images belonging to the same class (called the positive examples) and as far as possible from images belonging to a different class (called the negative examples). In other words, the triplet loss function

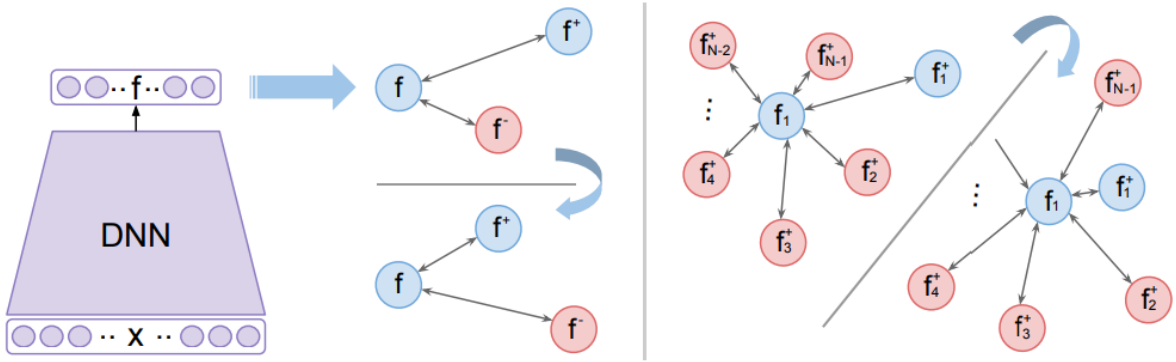


Figure 3.10: Comparison of deep metric learning with (left) triplet loss and (right) (N+1)-tuplet loss. Credits : Illustration from [Sohn, 2016]

forces the network to simultaneously reduce the distance between the deep features of the same class and increase the distance between deep features of different classes. This is achieved by minimizing the following loss function :

$$\mathcal{L}_{triplet} = \sum_{i=1}^{N} \max(d(f(x_i^a), f(x_i^p)) - d(f(x_i^a), f(x_i^n)) + \epsilon, 0) \tag{3.19}$$

where d is a distance metric, ϵ is a margin enforced between positive and negative pairs, and $f(x)$ represents the deep learned feature of an image x.

Npairs Loss : Although the success of deep metric learning frameworks based on triplet loss, these models still suffer from slow convergence. To overcome this problem, a new objective function called N-pair loss was proposed in [Sohn, 2016]. The N-pairs Loss generalizes the triplet loss by using more than one negative example and reduces the computational burden through a practical batch construction method. The authors propose two variations of the triplet loss function : (N+1)-tuplet loss and the N-pairs loss. Figure 3.10 illustrates the difference between the triplet loss and the (N+1)-tuplet loss. As can be seen, the triplet loss pulls a positive example while pushing one negative example at a time. In contrast, (N+1)-tuplet loss pushes away N-1 negative examples simultaneously based on their similarity to the input example. Consider an (N+1)-tuplet of training examples $\{x, x^+, x_1, \ldots, x_{N-1}\}$, where x^+ is a positive example to x and $\{x_i\}_{i=1}^{N-1}$ are negative. The (N+1)-tuplet loss is defined as follows :

$$\mathcal{L}_{(N+1)-tuplet} = \log\left(1 + \sum_{i=1}^{N-1} \exp\left(f^T(x)f(x_i) - f^T(x)f(x^+)\right)\right) \tag{3.20}$$

However, this $\mathcal{L}_{(N+1)-tuplet}$ loss is computationally expensive. For example, a batch size of M requires $M \times (N+1)$ examples to be passed through f at each update. To address this problem, authors introduce N-pair loss with a practical batch construction to avoid the excessive computational burden. Figure 3.11 illustrates this batch construction process. The multi-class N-pair loss is formulated as follows :

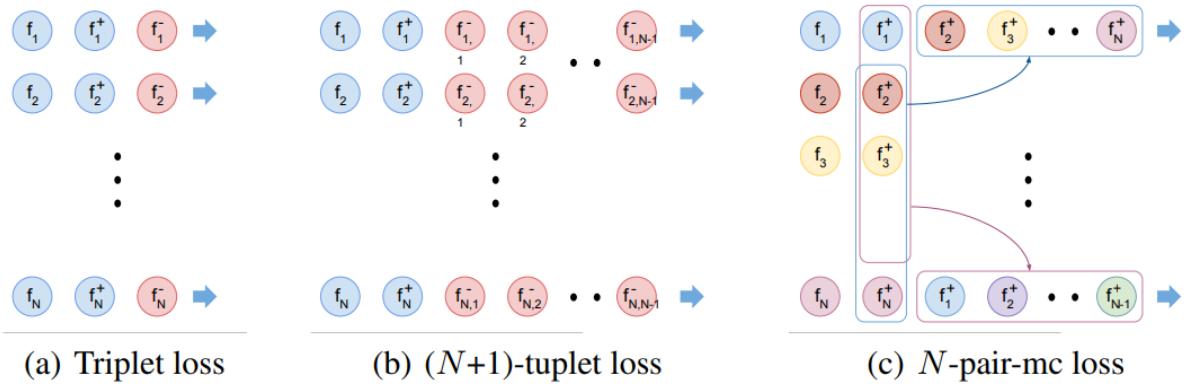


Figure 3.11: Triplet loss, (N+1)-tuplet loss, and multi-class N-pair loss with training batch construction. Credits : Illustration from [Sohn, 2016]

$$\mathcal{L}_{N-pair} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\sum_{j\neq i}\exp\left(f^T(x_i)f(x_j^+) - f^T(x_i)f(x_i^+)\right)\right) \tag{3.21}$$

By proposing $\mathcal{L}_{N-pair}$, the authors significantly improve the triplet loss by pushing away multiple negative examples jointly at each update.

Other popular methods use a hybrid objective function to improve the discriminative power of the CNN models by combining the discriminative loss with the softmax loss to train deep networks :

Center loss : In [Wen et al., 2016], the authors joined the softmax loss $\mathcal{L}_s$ with a new discriminative loss named center loss $\mathcal{L}_c$ to learn the most discriminative deep features. More specifically, the proposed center loss compute simultaneously c_{y_i} the center for deep features of each class and penalizes the distances between the deep features $\{f_i|i = 1, .., m\}$ and their corresponding class center c_{y_i}. The goal is to learn deep features with an inter-class dispensation and intra-class compactness as much as possible. The proposed center loss is formulated as follows :

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{m}\|f_i - c_{y_i}\|_2^2 \tag{3.22}$$

where m is the size of the mini-batch.

Island loss : is a new loss function proposed in [Cai et al., 2018] that builds upon the work of [Wen et al., 2016]. Unlike the center loss, which only minimizes intra-class variations, the island loss aims to reduce intra-class variations and increase inter-class differences simultaneously. It is defined as the sum of the center loss $\mathcal{L}_c$ and the pairwise distances between class centers in the feature space :

$$\mathcal{L}_{IL} = \mathcal{L}_c + \lambda\sum_{c_j\in N}\sum_{\substack{c_k\in N\\ c_k\neq c_j}}\left(\frac{c_j.c_k}{\|c_j\|_2.\|c_k\|_2}+1\right) \tag{3.23}$$

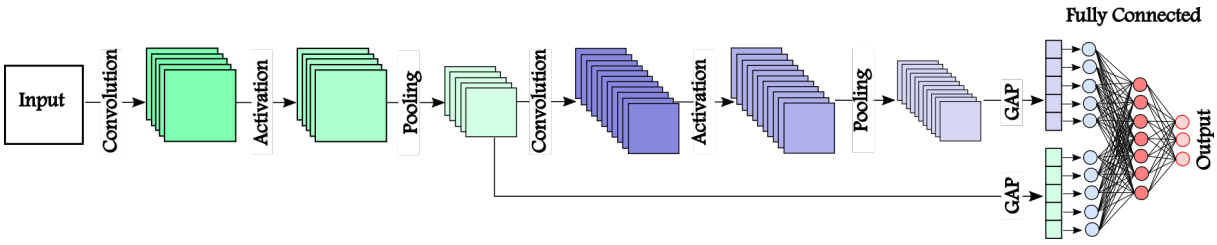


Figure 3.12: Illustration of the Texture Convolutional Neural Networks (TCNN) proposed in [Andrearczyk and Whelan, 2016].

where N is the set of computed class centers. The first term $\mathcal{L}_c$ penalizes the distance between the sample and its corresponding center, and the second term penalizes the similarity between class centers. By minimizing the $\mathcal{L}_{IL}$, the samples of the same class will get closer to each other, and those of different classes will be pushed apart.

3.3.3.C Texture CNNs

During the past few years, many CNN-based texture analysis methods have been proposed. One of the motivations behind using CNNs in texture analysis is the high similarity between the basic CNN operations (convolution, activation, and aggregation) and traditional filter bank methods, with the main difference being that CNN filters are learned directly from data rather than being predetermined by handcrafted filters [Liu et al., 2019]. Another reason is the architecture of classical CNNs, which tends to increase the abstraction level of the representation with depth, as shown in Figure 3.9. The first layers are designed to learn low-level features, such as edges and curves, which characterize the texture information, while the deeper layers are trained to capture more complex and high-level patterns, such as overall shape information. Consequently, as a result of successive CNN operations, the fine texture-related image details tend to disappear in the top layers of the network.

As shown in Figures 3.8, for a standard CNN, the output of the last convolutional layer is reshaped into a single feature vector representing the global spatial information. However, this spatial information is necessary for analyzing the global shapes of objects, but it is not of great importance for analyzing textures.

To overcome this problem, several methods are proposed. In [Andrearczyk and Whelan, 2016], authors proposed a simple solution by combining the global average pooled vector from the outputs of multiple convolutional layers with that of the last convolutional layer using a concatenation layer and then feeding them to the fully connected (classification) layers. The proposed network, called Texture CNN (TCNN), is illustrated in Figure 3.12.

In [Lin et al., 2017], authors proposed to replace the fully connected layers with an order-less bilinear pooling layer, as shown in Figure 3.13. This network, called BCNN, achieved slightly better results than the existing CNN-based texture classification methods. However, the representational power of bilinear

features comes at the expense of very high dimensional feature representations, which leads to significant computational burdens.

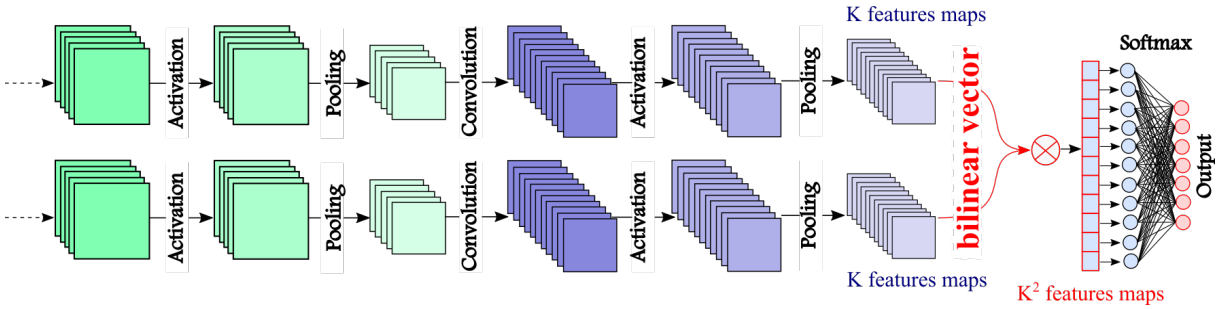


Figure 3.13: Illustration of the Bilinear Convolutional Neural Networks (BCNN) proposed in [Lin et al., 2017].

To reduce the dimensionality of bilinear features, Gao et al. suggest a compact bilinear pooling [Gao et al., 2016], as illustrated in Figure 3.14. The proposed method uses specific random projection techniques, such as Random Maclaurin Projection and Tensor Sketch Projection, to reduce the dimensionality of bilinear representations by 90% while maintaining similar performance to the BCNN network.

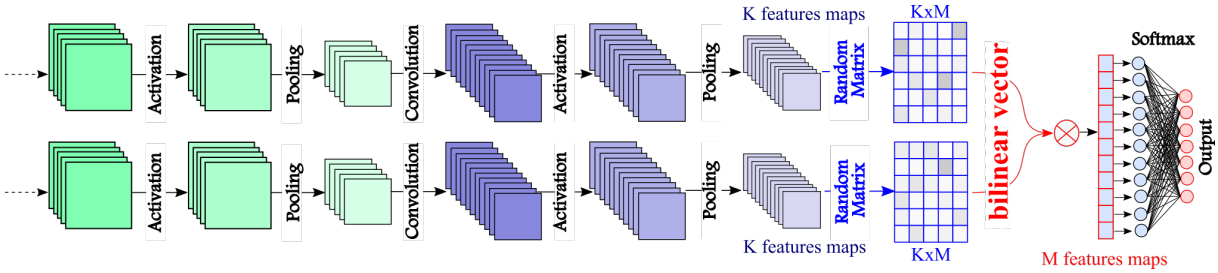


Figure 3.14: Bilinear Convolutional Neural Networks with a compact bilinear pooling to reduce the feature dimensionality [Gao et al., 2016].

In [Dai et al., 2017], Dai *et al.* introduced an effective fusion network called FASON (First And Second Order information fusion Network) that combines the ideas of TCNN [Andrearczyk and Whelan, 2016] and Compact BCNN [Gao et al., 2016]. As illustrated in Figure 3.15, FASON combines first and second-order information from multiple convolutional layers and concatenates them to form a single representation vector, which is then connected to the fully connected classification layer.

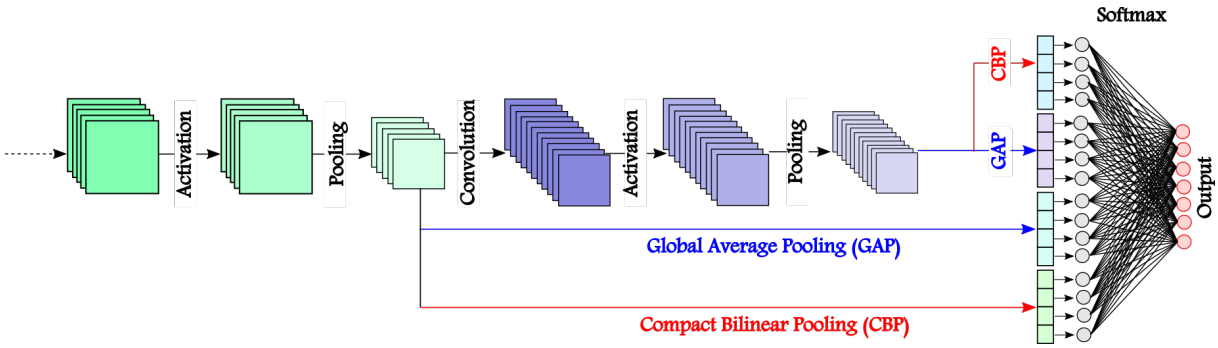


Figure 3.15: Illustration of First and Second Order Information Fusion Network proposed in [Dai et al., 2017].

4

Discriminative Regularized Auto-Encoder for Early Knee Osteoarthritis Detection

Contents

4.1 Introduction

In this chapter, we introduce a new approach for early detection of knee OA using X-ray images. We mainly focus on the feature learning step as a crucial component on the classification system in order to learn and extract the most useful discriminative features from pixel intensities. For that, we introduce a Discriminative Regularized Auto-Encoder (DRAE). More specifically, a penalty term, called discriminative loss, is combined with the standard Auto-Encoder training criterion which aims to force the learned representation to contain discriminative information.

The proposed approach performs a series of consecutive steps. Firstly, a semi-automatic segmentation method is used to localize the knee joint. Then, five Regions Of Interest (ROI) of size 32×32 pixels (Figure 4.5) are extracted from the bone under the tibial plateau, representing the lateral, central, and medial parts of the knee. Then, the proposed DRAE is applied to extract high-level features of each ROI. Next, the obtained features are used to train the classifier to differentiate between normal case (KL grade 0) and early OA (KL grades 1 or 2). Once the classifier is trained, we feed it with the test dataset to predict their corresponding labels. Finally, we qualitatively and quantitatively evaluate the obtained results. These steps were summarized in the pipeline illustrated in Figure 4.1.

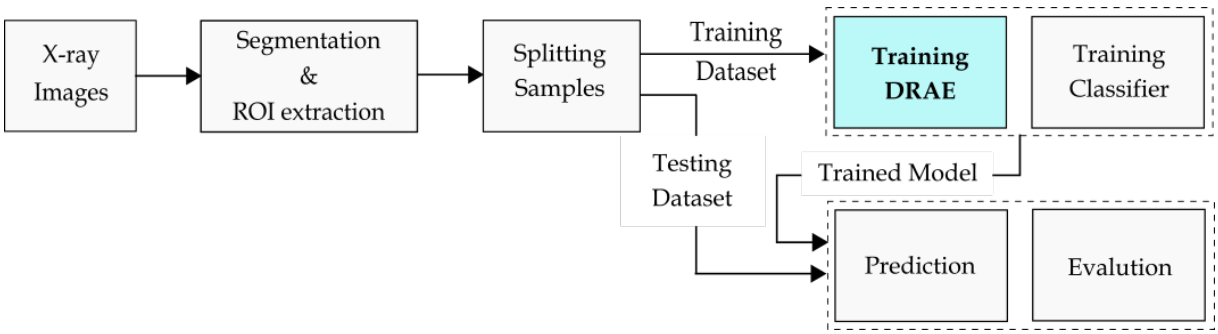


Figure 4.1: The pipeline of the proposed method

4.2 Discriminative Regularized Auto-Encoder (DRAE)

Traditional Auto-Encoders (AEs) have successfully shown their ability of learning high level features from raw data in various applications [Litjens et al., 2017, Shin et al., 2012, Xu et al., 2015]. However, extracting meaningful patterns from X-ray images can be a challenging task due to the complex nature of the images and the high similarity between radiographs of knee OA cases and healthy subjects. As explained in section 3.3.1, AEs aims to reduce the dimension of the input by keeping only the most relevant information. However, when the input images are very similar, the distributions of the learned features of the two classes overlap (see Figure 4.2).

To deal with this issue, we propose a novel AE-based architecture named Discriminative Regularized

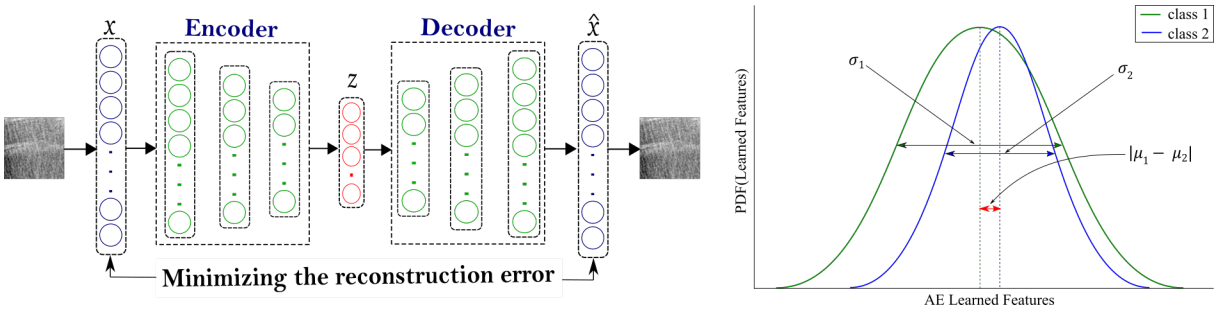


Figure 4.2: Flowchart of the proposed DRAE. The discriminant penalty is adopted in the hidden layer to maximize the separability between feature classes

Auto-encoder (DRAE), illustrated in Figure 4.3. In addition to the standard reconstruction objective, DRAE forces the network to capture discriminative properties that maximize the distance between feature classes. Hence, DRAE has two main goals : force the network to capture discriminative properties that *(i)* minimize the intra-class distance and *(ii)* maximize the inter-class distance. To this end, an additional term called discriminative penalty is added to the initial cost function. The overall objective function is finally defined in this study as a combination of three terms : the reconstruction error, the discriminative penalty, and a L_2 regularization term, as depicted by Eq. 4.1 :

$$J_{DRAE} = J_{AE} + \lambda\Omega_{disc} + \beta\Omega_{weights} \tag{4.1}$$

λ and β are respectively the weights of the discriminative and L_2 regularization terms.

J_{AE} is the Mean Square Error (MSE), which is the cost function of the traditional AE (Eq. 3.14). MSE measures how the reconstructed input $\hat{x}$ is close to the original input x, by finding the optimal parameters that minimize the error of reconstruction. In addition, it allows the hidden units to encode the structures that can be used as good representations of the input data.

Ω_{disc} represents the discriminative penalty. Inspired from the well-known Fisher Linear Discriminant

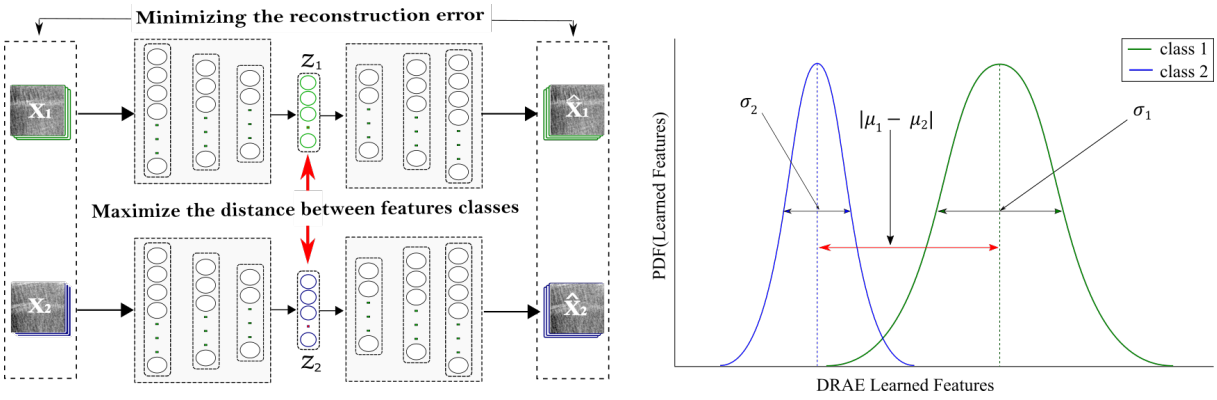


Figure 4.3: Flowchart of the proposed DRAE. The discriminant penalty is adopted in the hidden layer to maximize the separability between feature classes

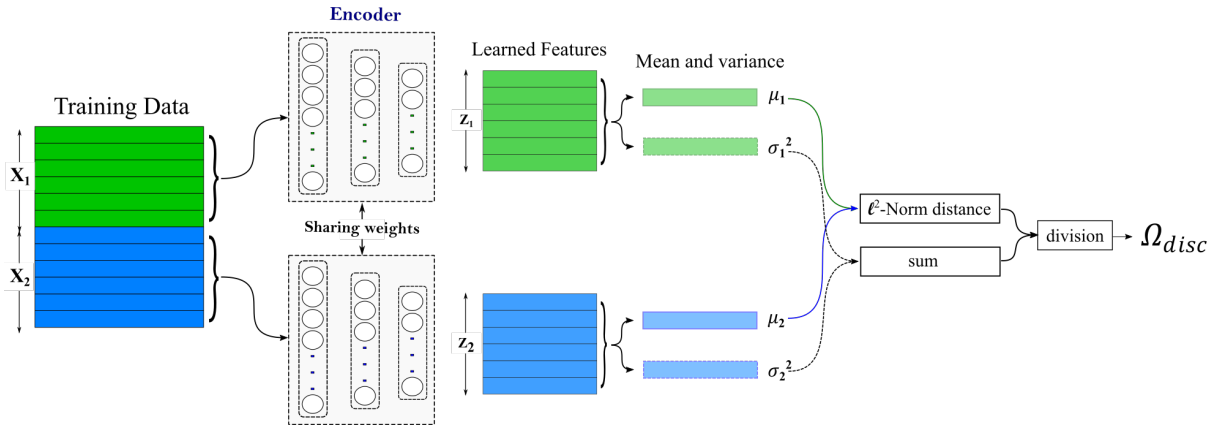


Figure 4.4: Visualized steps to calculate the the discriminative penalty Ω_{disc}.

Analysis (FLDA) algorithm [Fisher, 1936, Mika et al., 1999], Ω_{disc} attempts to encourage classes separability by maximizing the distance between the means μ_1 and μ_2 of the encoding feature sets (Z_1 and Z_2) of each class and minimizing their variances σ_1^2 and σ_2^2. Ω_{disc} is defined as follows :

$$\Omega_{disc} = \frac{\sigma_1^2 + \sigma_2^2}{|\mu_1 - \mu_2|^2} \tag{4.2}$$

where,

$$\mu_i = \frac{1}{m^{(i)}} \sum_{j=1}^{m^{(i)}} z_j^{(i)} \quad ; \quad \sigma_i^2 = \frac{1}{m^{(i)}} \sum_{j=1}^{m^{(i)}} (z_j^{(i)} - \mu_i)^2 \tag{4.3}$$

μ_i and σ_i^2 are respectively the mean and variance of the set of learned features, $Z_i = \{z_j^{(i)} | j = 1, ..., m^{(i)}\}$ of the class i. $m^{(i)}$ is the number of samples in class i. According to Eq. 4.2, Ω_{disc} takes a large value when the distributions of the two classes are close and become small as they move away from each other. Figure 4.4 visualize the steps of computing the discriminative penalty Ω_{disc}.

$\Omega_{weights}$ is a regularization term of the weights called L_2 regularization and is defined by $\Omega_{weights} = \frac{1}{2}\sum_{l=1}^{L} \|W^l\|_2^2$, where L is the number of hidden layers. $\Omega_{weights}$ is used to force the weights to become small to avoid overfitting [Moody, 1991] and to make the model more stable [Tartaglione et al., 2018].

The optimization of the J_{DRAE} objective function (Eq. 4.1) with respect to the network parameters θ is carried out using the full-batch gradient descent algorithm to find an optimal solution [LeCun et al., 2012]. Doing so leads to learning the complex patterns across data variations and extracting the feature vectors that will be fed to the classifier.

Algorithm 1 describes the learning process of the proposed model. First, the training data are split into two sets X_1 and X_2 corresponding to the input data of both classes (normal and minimal OA). Second, the set of the network parameters θ is initialized for training. Third, the set of hidden representations Z_1 and Z_2, and their corresponding means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), and the reconstructed inputs $\hat{X}_1$ and $\hat{X}_2$ are computed as mentioned in section 3.3.1. The back-propagation algorithm [LeCun et al.,

Algorithm 1 Learning Algorithm of the DRAE

Input : $\{X_1, X_2, \alpha, \lambda, \beta\}$
Output : final learned features and network parameters
repeat
  Compute the sets of codes Z_1 and Z_2
  Compute mean and variance of Z_1 and Z_2
  Compute reconstructed inputs sets $\hat{X}_1$ and $\hat{X}_2$
  Update θ *(with 1-step of gradient descent)*
until *convergence*

2012] is used to find an optimal solution by computing the average gradient of the objective function, $\nabla J_{DRAE}(\theta)$, over all the training samples (full-batch gradient descent). Once the gradient is computed, the set of parameters θ is updated according to the following rule :

$$\theta = \theta - \alpha \nabla J_{DRAE}(\theta) \tag{4.4}$$

where α representing the learning rate. This process is repeated until the model converges.

4.3 Datasets and Experimental setups

4.3.1 Datasets

The database used in this study was obtained from the public baseline dataset Osteoarthritis Initiative (OAI) described in section 2.2.1. The OAI dataset is composed of bilateral fixed flexion knee radiographs of both men and women patients aged between 45 and 79 years old. Each knee joint is associated with its KL grade.

Analysis of the bone microarchitecture in OA date back more than 30 years and has provided clear indications that changes in the periarticular bone occur very early in OA development [Radin and Rose, 1986, Lynch et al., 1991a, Lynch et al., 1991b, Messent et al., 2005]. As we are mainly interested in early knee OA detection, different ROIs under the tibial plateau were extracted using the segmentation method proposed in [Janvier et al., 2017], which is a semi-automatic method, due to the large variations of the acquisition parameters (resolution, contrast, exposure, ...). The main steps are summarized in the following. Firstly, anatomical markers, i.e. the tibial spines, and the lateral and medial extremities of the tibia are manually marked (see red dots in Figure 4.5.b). Then, the tibial subchondral bone plate is automatically detected as the brightest path going through these anatomical markers. Finally, five square ROIs (R_1 to R_5) of size 32×32 pixels are extracted under the inferior border of the cortical plates and representing the medial (R_1, R_2), central (R_3) and lateral (R_4, R_5) parts of the knee, as shown in Figure 4.5.b.

The KL scoring system is a clinical grade used to score knee OA in the clinical routine. As there is no

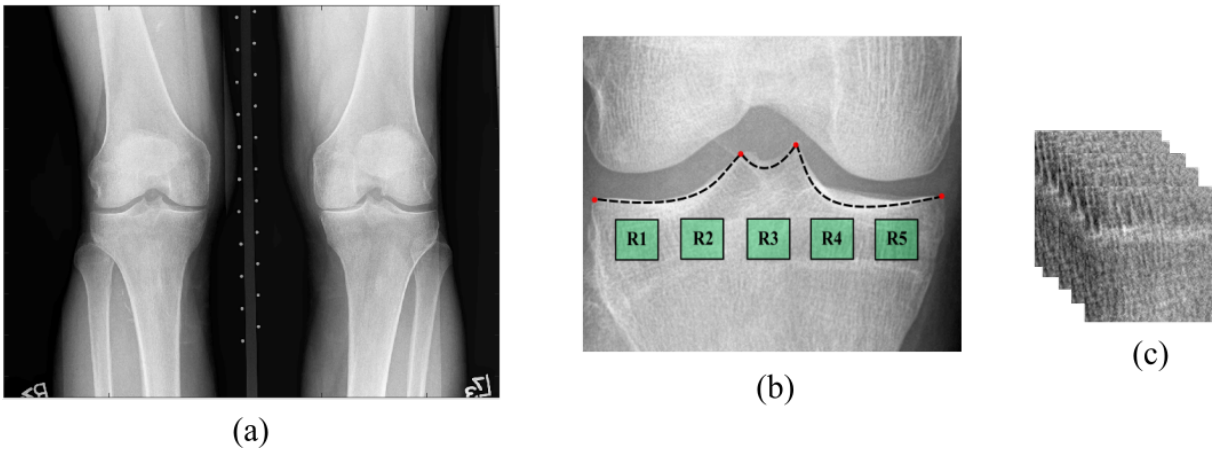


Figure 4.5: Selected ROIs using a semi-automatic algorithm. (a) a typical knee X-ray image, (b) a knee joint with the selected ROIs, the dashed line represents the tibial edge and (c) a set of extracted ROIs.

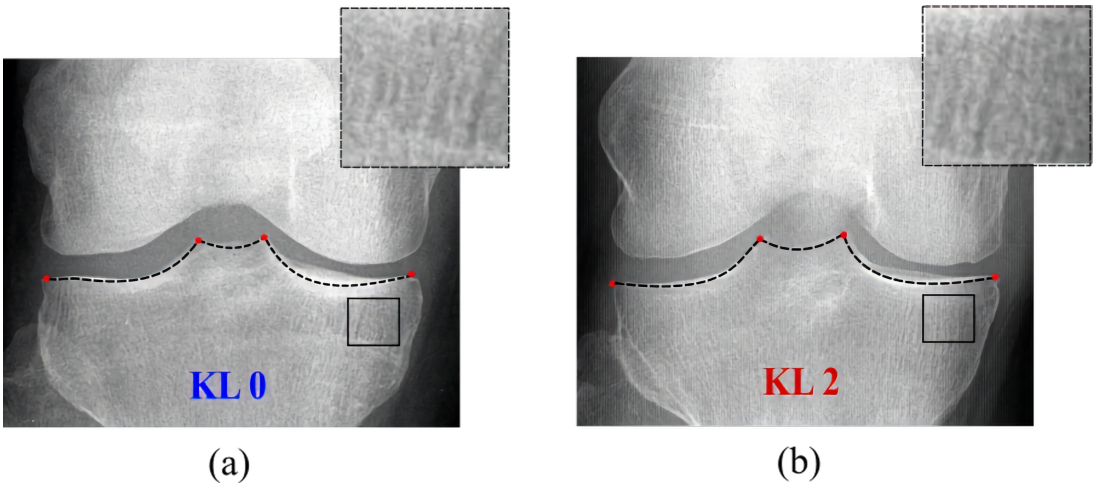


Figure 4.6: Typical knee joint area and ROIs extracted from knee radiographs of : (*a*) a healthy subject (KL grade 0), (*b*) an OA case (KL grade 2).

scoring system based on texture analysis for knee OA, we considered the existing KL score given by the attribute "V00XRKL" in the OAI study to classify the subjects as normal or abnormal. In this study, our objective is to distinguish between the definite absence (KL grade 0) and the early presence of OA (KL-G1 and KL-G2). It is a challenging task due to the high degree of similarity between both cases as shown in Fig 4.6. 3900 knee joint images from the OAI baseline dataset were used. 1300 images from each grade (0, 1, and 2) to avoid any statistical bias. Moreover, only the Computed Radiography (CR) modality was considered to avoid digitizing artifacts. The proposed DRAE is binary, to avoid ambiguous effects that might be introduced by the doubtful KL-G1 class, our model is first validated and tested using only the data from KL-G0 and KL-G2 classes.

Figure 4.7 aims at showing that there is an overlap between the distributions of ROIs extracted from the two random images. This overlap explains the high similarity between the samples of the two classes.

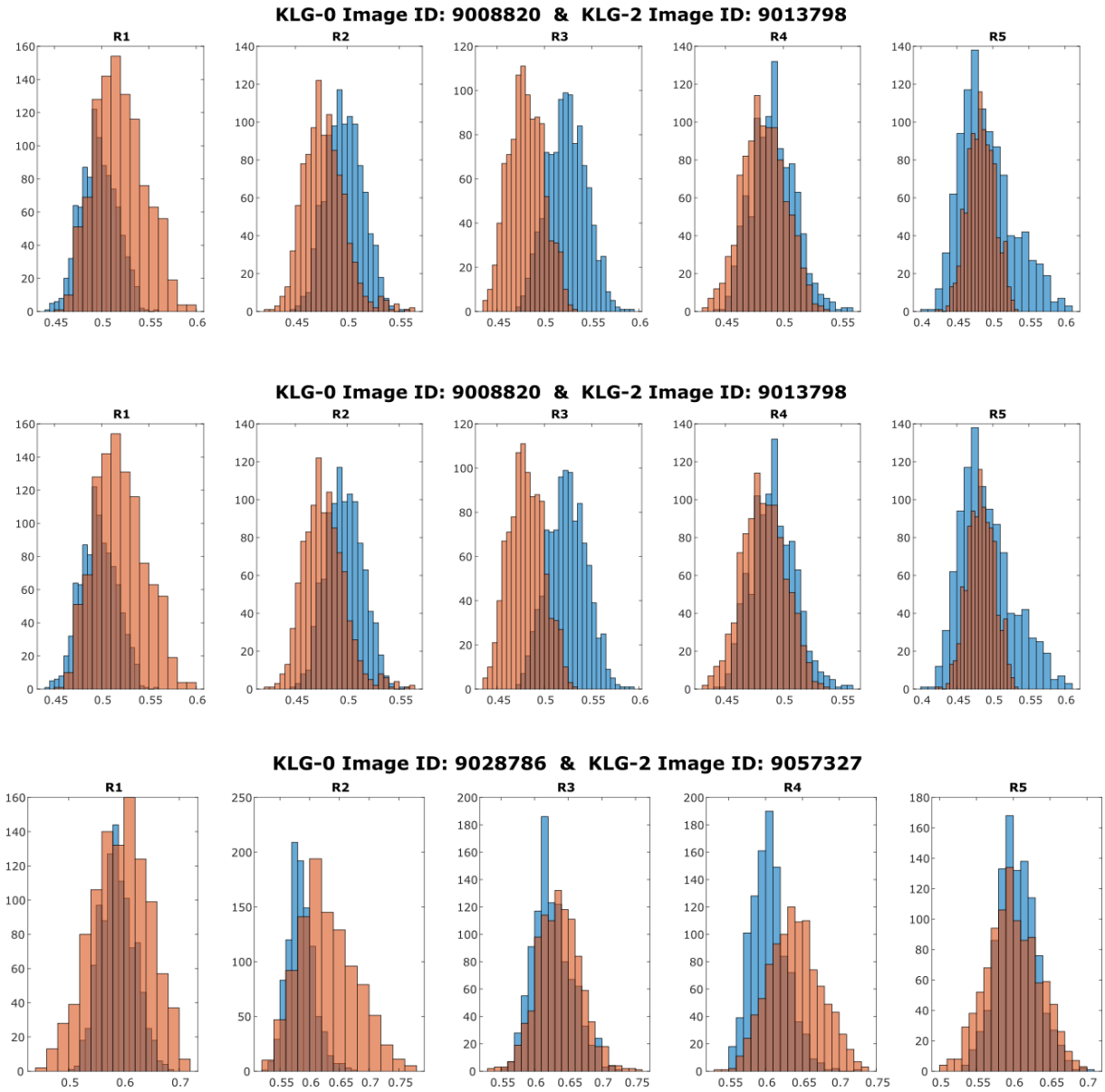


Figure 4.7: Different examples of ROIs histograms from different randomly taking images.

4.3.2 Experimental setups

4.3.2.A Training phase

As mentioned previously, the input size of our DRAE is an ROI of size 32×32, which is flattened and directly connected to the first layer of the DRAE network. In this model, we use tied weights ($W_d^l = (W_e^l)^T$), the tangent hyperbolic ($tanh$), and the sigmoid activation functions for both encoder and decoder, respectively. The Root Mean Square Error (RMSE) was used to compute the reconstruction loss. Three hidden layers were used. The number of the hidden units K_i of each layer $\{l_i | i = 1, 2, 3\}$ are $K_1 = K$, $K_2 = \frac{3}{4}K$, and $K_3 = \frac{1}{2}K$. The DRAE model was trained and optimized end-to-end using a full-batch SGD optimizer.

Once the learning procedure is completed, each resulting feature vector z_i is coupled with its corresponding label y_i as $\{(z_i, y_i) | i = 1, ..., N\}$ and then fed to the classifier to distinguish between the normal knee and OA one.

4.3.2.B Evaluation metrics

The evaluation was conducted by applying a 10-fold cross-validation. The data was divided into 10 equal folds and the learning process was repeat 10 times. In each step, the DRAE and the classifier were trained and validated on 9 folds, 8 folds for training, and one fold for validation. Then, the model was tested on the remaining portion. The classification performance on the validation set was used to select the best set of hyperparameters $\{K, \lambda, \beta\}$. The final performance was computed as the average of the obtained values. The classification performance was performed using 4 metrics :

- Accuracy (Acc) : is the percentage of predictions that match exactly the ground-truth, computed as :

$$\text{A}ccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

- Precision (Pr) : is the fraction of true positives among predicted positives, computed as :

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

- Recall (Re) : is the fraction of a total number of true positives retrieved, computed as :

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

- F1-score (F1) : also called F-measure, is the harmonic mean of precision and recall, it is very useful

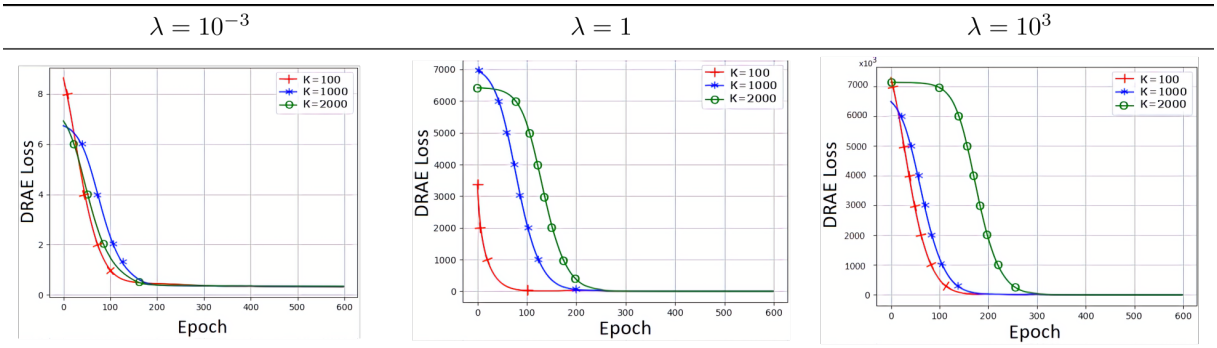


Figure 4.8: Obtained convergence test curves by increasing K and varying the value of the penalty parameter (λ).

in describing model performance and in comparing models, and is defined as follows :

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.8}$$

where, TP and TN correspond respectively to True Positive and True Negative samples, while FP and FN represent False Positive and False Negative errors.

For the technical details, Tensorflow library [Abadi et al., 2016] was employed to implement our model. The tests were executed using a workstation with CPU Intel(R) Core(TM) i7-7700 @ 3.4 GHz with 16 GB of RAM and a GTX 1050 Ti NVIDIA Graphics card with 4 GB memory.

4.4 Experimental Results

In this section, our proposed DRAE is evaluated on data from the OAI database. Firstly, the convergence of the proposed model is analyzed and the learned representation is visualized. Then, the influence of the hidden unit size and the weight of the discriminative penalty on classification performance are studied. Finally, several classifiers are tested and the classification performance of the proposed DRAE is compared to both traditional AE [Bourlard and Kamp, 1988] and Sparse Auto-Encoder (SAE) [Ng et al., 2011] as well as to standard state-of-the-art deep learning-based methods : ResNet-101 [He et al., 2016] and DenseNet-121 [Huang et al., 2017].

4.4.1 Analysis of the convergence of the model

Herein, we examine how the proposed network training strategy behaves by increasing the number of hidden units and varying the penalty parameter values. Figure 4.8 reports the model performance over epochs obtained by varying the penalty values and increasing the capacity of the model in terms of

its number of neurons per layer. To visualize the stability of the model, we considered three different values for $\lambda = \{10^{-3}, 1, 10^{3}\}$ and three configurations of $K = \{100, 1000, 2000\}$. As can be seen, due to the increase in the number of parameters to be optimized, the higher K, the slower the convergence. Results show that even if the speed of the convergence differs in the different considered configurations, the training process in DRAE is guaranteed to converge in a finite number of epochs up to 300, which ensures the global stability of the model.

4.4.2 Visualization of the Learned Features

This section aims at verifying if the learned representation using our model could maximize the class separability and capture the useful discriminative features for the classifier. To this end, we compared the obtained representations learned by the proposed DRAE and those learned by the traditional AE using 2D scatter plots of t-distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten and Hinton, 2008]. More specifically, from the training samples of each ROI (R_1, R_2, R_3, R_4, R_5), a set of feature vectors was extracted and the dimension of each vector was reduced to 2 using t-SNE. The DRAE was tested using 100 and 1000 hidden units to see the effect of the size of the feature vector. Results are illustrated in Figure 4.9. The first column shows the original input space (raw data). As can be seen, the two classes significantly overlap. The next two columns show the feature vectors obtained by applying AE with 100 and 1000 hidden units, respectively. In this case, the learned features by the traditional AE increase the separation between the two classes in the transformed space, but not enough to identify OA cases. The last two columns show the learned features obtained when DRAE is applied with 100 and 1000 hidden units, respectively. Results show that DRAE learned two discriminant representations and increases the separation of features from the two classes. The aggregation of features within the same class was also reinforced, particularly when the number of hidden units is small. These results suggest that DRAE works well in learning discriminative features and thus will enable better classification of knee OA.

4.4.3 Hidden units (K) and the weight of the discriminative penalty (λ) effects

In order to study the influence of the number of the hidden units, K and the weights of the discriminative penalty, λ on the classification performance, these hyper-parameters were tuned from two sets : $K = \{100, 200, 400, 800, 1000, 1600\}$ and $\lambda = \{1 \times 10^{e} | e = -6, -5, ..., 3\}$. Figure 4.10 illustrates the evolution of the accuracy depending on K and λ of the proposed DRAE coupled to the Support Vector Machine classifier with the Radial Basis Function (SVM-RBF) for ROI-R2. As can be seen, the performance of the model depends mainly on the weights of the discriminative penalty, λ. A good accuracy can be obtained only if $\lambda \leq 0.1$. For $\lambda > 0.1$, the accuracy decreases remarkably. Furthermore, for $\lambda \leq 0.1$, the model

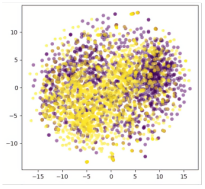
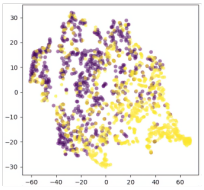
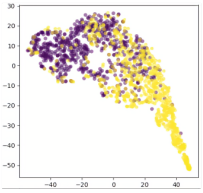
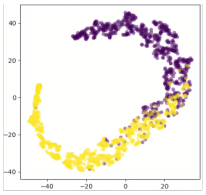
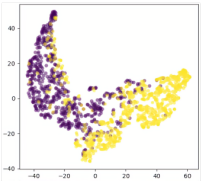
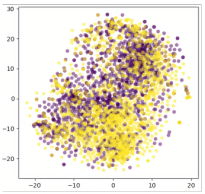
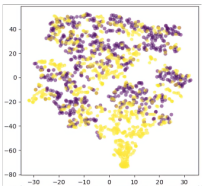
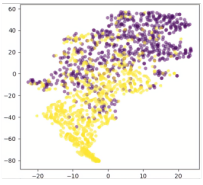
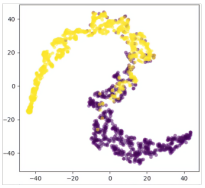
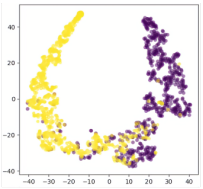
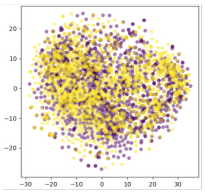
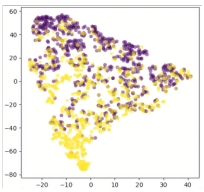
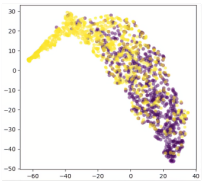
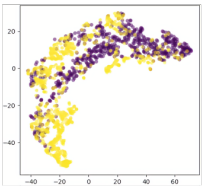
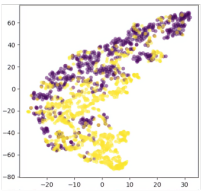
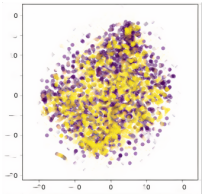
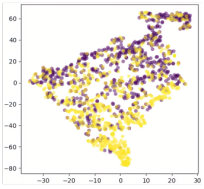
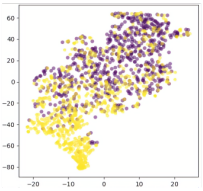
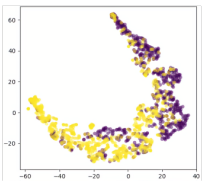
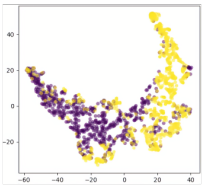
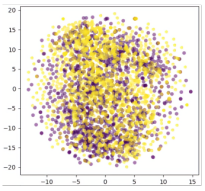
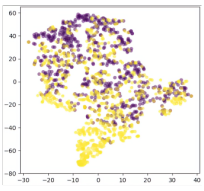
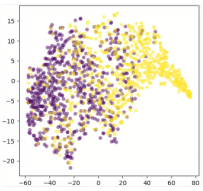
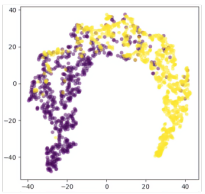
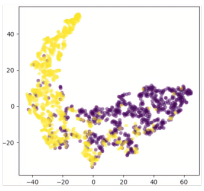
ROIs	Original data	AE with K=100	AE with K=1000	DRAE with K=100	DRAE with K=1000
R1					
R2					
R3					
R4					
R5					

Figure 4.9: Obtained t-SNE scatter plots for each ROI using raw data and learned features by AE and DRAE models for $K = \{100, 1000\}$.

achieves good performance when the value of K is less than the input size (i.e. $K \leq 1000$). Thus, high performance is achieved by the proposed model when the value of λ is less than 0.1 and the value of K is less than the input size, I.

4.4.4 Classification performance using different classifiers

4.4.4.A KL-0 vs. KL-2 Classification Results

We compared five commonly used learning algorithms : SVM-RBF [Chang and Lin, 2011], Linear Discriminant Analysis (LDA) [Fisher, 1936], Softmax Classifier (SMC) [Bishop and Nasrabadi, 2006], and the K-Nearest Neighbours (KNN) [Goldberger et al., 2004] with the Euclidean distance. These algorithms were performed in the binary classification case. The purpose of this experiment is to select the best classifier that achieves the highest classification performance. Table 4.1 summarizes the results obtained for the 5 ROIs. To have a single decision per image, a max-voting strategy was applied as follows :

$$\hat{y}_i = mode\{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}, \hat{y}_i^{(4)}, \hat{y}_i^{(5)}\} \tag{4.9}$$

where $\hat{y}_i^{(j)}$ is the predicted label of the ROI R_j extracted from the i^{th} image. *mode* represents the most frequent value. The max-voting score was computed using the obtained labels $\hat{y}_i$ and the true labels y_i.

According to Table 4.1, the highest performance was obtained using the SVM-RBF classifier. It achieved a max-voting accuracy of 82.53% compared to KNN (78.15%), SMC (78.69%), and LDA (72.69%). In terms of the F1-score, the highest value (83.48%) was also obtained using the SVM classifier, corresponding to 88.23% in precision and 79.22% in the recall. These results show that the SVM-RBF classifier, yields in general better performance compared to the other learning algorithms. Therefore, it

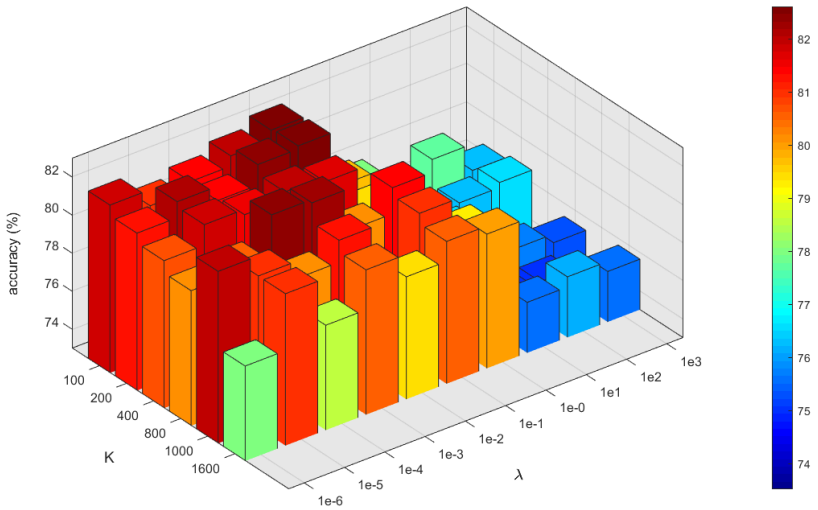


Figure 4.10: Obtained performance using DRAE+SVM with different values of K and λ for ROI-R2.

Table 4.1: CLASSIFICATION PERFORMANCE (KL-G0 VS. KL-G2) FOR EACH ROI USING DIFFERENT CLASSIFIERS

Classifiers	ROIs	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM-RBF	1	80.07	83.61	78.08	80.75
	2	81.73	86.15	79.17	82.51
	3	74.26	77.07	73.07	75.02
	4	71.50	78.30	68.97	73.34
	5	77.23	80.84	75.37	78.01
	Max-voting	**82.53**	**88.23**	**79.22**	**83.48**
LDA	1	71.23	72.84	70.57	71.69
	2	65.15	67.69	64.33	65.97
	3	62.11	62.30	62.06	62.18
	4	61.07	63.69	60.51	62.06
	5	61.57	63.00	61.29	62.13
	Max-voting	72.69	75.53	71.45	73.43
SMC	1	74.96	74.23	75.33	74.78
	2	77.50	79.84	76.27	78.01
	3	66.46	67.76	66.06	66.90
	4	65.57	67.53	65.05	66.27
	5	72.84	75.07	71.90	73.45
	Max-voting	78.69	83.53	76.15	79.67
KNN	1	73.03	74.53	72.35	73.43
	2	76.50	77.15	76.15	76.65
	3	71.69	72.76	71.30	72.02
	4	68.88	71	68.12	69.53
	5	72.26	72.69	72.08	72.38
	Max-voting	78.15	82.00	76.15	78.96

was retained for the next experiments.

4.4.4.B KL-G1 Classification Results

In this section, we have evaluated the performance of the proposed DRAE using KL-G1. As the proposed DRAE is a binary network affording only two classes, we performed a classification task considering (KL-G0 vs. KL-G1) and (KL-G1 vs. KL-G2). Results are shown in Tables 4.2 and 4.3. These results prove that our proposed DRAE is also competitive for the classification of KL-G1 patients. We recall that KL-G1 has been reported in the literature as a doubtful class. Analysis of the results in Table 4.2 shows that accuracy of 69.83% and an F-score of 70.95% were achieved for KL-G0 vs. KL-G1. Considering KL-G1 vs. KL-G2, a better performance was reached with an accuracy of 77.05% and an F-score of 77.21% as shown in Table 4.3. Overall, these results show that KL-G0 and KL-G1 classes are very similar making the classification task more challenging.

4.4.5 Comparison to other models

Our proposed DRAE was firstly compared to unsupervised learning-based auto-encoder models : classical Auto-Encoder (AE) and Sparse Auto-Encoder(SAE). Results are shown in Table 4.4 in terms of accuracy, precision, recall, and F1-score obtained using the SVM-RBF classifier. As can be seen, the accuracy and F1-score obtained using the proposed DRAE are higher than those obtained using the classical AE and SAE models. The max-voting score shows that the proposed DRAE reaches an accuracy of 82.53% compared to 81.11% and 80.26% achieved by the SAE and AE models, respectively. In addition, DRAE obtains a F1-score of 83.48% compared to 81.43% and 80.87% reached by SAE and AE, respectively.

The proposed DRAE was also compared to standard state-of-the-art deep learning-based models : ResNet-101 [He et al., 2016], and DenseNet-121 [Huang et al., 2017]. To this end : first, we use pretrained weights as a starting point. Then, we replace the final classification layer with a new layer adapted to the new data set. Finally, we retrain the pretrained network to learn the new OA detection task. Fine-tuning a network is usually much faster and easier than training a network from scratch with randomly initialized weights [Tajbakhsh et al., 2016]. Results of the max-voting are reported in Table 4.5. As can be seen, DRAE provides better performance than the well-known convolutional neural networks, ResNet-101 and DenseNet-121. ResNet-101 achieves an accuracy of 59.25% and an F1-score of 69.43%, while

Table 4.2: (KL-G0 VS. KL-G1) CLASSIFICATION PERFORMANCE USING DRAE AND THE SVM-RBF CLASSIFIER

ROIs	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	67.51 ± 0.91	74.40 ± 2.12	65.40 ± 0.92	69.59 ± 1.03
2	68.15 ± 1.36	71.50 ± 1.56	67.04 ± 1.57	69.18 ± 1.17
3	62.22 ± 1.11	63.59 ± 2.39	61.92 ± 1.15	62.72 ± 1.35
4	59.19 ± 0.92	63.07 ± 2.84	58.52 ± 0.70	60.70 ± 1.52
5	66.00 ± 0.98	68.09 ± 1.38	65.39 ± 1.29	66.70 ± 0.76
Max-voting	$\mathbf{69.83 \pm 1.02}$	$\mathbf{73.68 \pm 1.07}$	$\mathbf{68.43 \pm 1.30}$	$\mathbf{70.95 \pm 0.78}$

Table 4.3: (KL-G1 VS. KL-G2) CLASSIFICATION PERFORMANCE USING DRAE AND THE SVM-RBF CLASSIFIER

ROIs	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	74.37 ± 0.43	75.29 ± 1.72	73.94 ± 0.59	74.59 ± 0.69
2	75.69 ± 1.68	73.20 ± 2.00	77.10 ± 2.44	75.07 ± 1.57
3	63.35 ± 1.86	66.17 ± 2.26	62.70 ± 2.04	64.36 ± 1.62
4	62.54 ± 1.76	63.33 ± 2.25	62.37 ± 1.83	62.83 ± 1.75
5	65.52 ± 1.18	69.10 ± 2.95	64.53 ± 1.47	66.70 ± 1.33
Max-voting	$\mathbf{77.05 \pm 0.37}$	$\mathbf{77.80 \pm 1.87}$	$\mathbf{76.67 \pm 0.91}$	$\mathbf{77.21 \pm 0.60}$

DenseNet-121 achieves 61.30% and 71.67% in terms of accuracy and F1-score, respectively. This proves that the proposed DRAE model can extract meaningful patterns to improve the overall classification, especially in this challenging task, where a high similarity exists between ROIs from early knee OA cases and healthy subjects. For comparison purposes, we have also tested the model of [Tiulpin et al., 2018] using our samples and the extracted ROIs of [Tiulpin et al., 2018]. Obtained results are shown in table 4.5. As can be seen, our method performs slightly better. This improvement can be explained by the fact that in [Tiulpin et al., 2018], the authors used a Siamese CNN, more suited for shape analysis, while we exploit texture information along with the regularized DRAE.

4.4.6 Summary of results

Obtained results for each ROI (see Table 4.4) show that the most distinguishing regions, R_1 and R_2 are located on the medial side of the tibial region, with an accuracy of 80.07% and 81.73%, respectively. The lowest rates were obtained with ROIs R_3 and R_4 located in the central-medial parts. A possible reason is that OA progression occurs mostly in the medial compartment due to more biomechanical load on this site of the knee [Ledingham et al., 1993, Iorio and Healy, 2003, Radin and Rose, 1986]. These results demonstrate that the localization of ROI plays a significant role in knee OA prediction. Our findings show that the placement of ROI could provide more discriminative information on changes in the bone microarchitecture due to OA and lead to a robust prediction. These preliminary results

Table 4.4: COMPARISON OF CLASSIFICATION PERFORMANCE FOR DRAE, SAE AND AE FOR EACH ROI

Methods	ROIs	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DRAE	1	80.07 ± 0.11	83.61 ± 0.32	78.08 ± 0.01	80.75 ± 0.14
	2	81.73 ± 1.03	86.15 ± 4.13	79.17 ± 0.71	82.51 ± 1.5
	3	74.26 ± 1.68	77.07 ± 1.74	73.07 ± 3.09	75.02 ± 0.80
	4	71.50 ± 0.59	78.30 ± 2.82	68.97 ± 1.67	73.34 ± 0.29
	5	77.23 ± 2.06	80.84 ± 4.02	75.37 ± 1.00	78.01 ± 2.41
	Max-voting	$\mathbf{82.53 \pm 0.87}$	$\mathbf{88.23 \pm 1.52}$	79.22 ± 0.43	$\mathbf{83.48 \pm 0.86}$
SAE	1	79.50 ± 0.38	83.15 ± 1.63	77.51 ± 1.38	80.23 ± 0.01
	2	80.88 ± 0.70	84.23 ± 2.71	78.95 ± 0.42	81.50 ± 1.04
	3	73.19 ± 0.70	80.38 ± 3.58	70.35 ± 2.14	75.03 ± 0.34
	4	71.65 ± 1.57	83.92 ± 5.11	67.37 ± 0.27	74.74 ± 2.20
	5	75.61 ± 1.84	81.46 ± 0.32	72.98 ± 2.55	76.98 ± 1.27
	Max-voting	81.11 ± 0.82	82.30 ± 2.83	$\mathbf{80.39 \pm 0.40}$	81.34 ± 1.18
AE	1	78.50 ± 1.68	82.69 ± 2.50	76.28 ± 1.16	79.36 ± 1.78
	2	79.30 ± 1.41	85.61 ± 4.46	76.02 ± 0.15	80.53 ± 1.88
	3	73.34 ± 2.01	78.46 ± 2.61	71.33 ± 3.61	74.72 ± 0.80
	4	71.53 ± 0.76	81.15 ± 2.50	68.10 ± 1.63	74.06 ± 0.07
	5	75.00 ± 1.84	82.23 ± 0.54	71.90 ± 2.11	76.69 ± 1.44
	Max-voting	80.26 ± 0.70	83.46 ± 0.76	78.45 ± 0.64	80.87 ± 0.70

Table 4.5: COMPARISON TO SATE OF THE ART DEEP LEARNING METHODS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-101	59.25 ± 0.47	78.26 ± 4.15	62.48 ± 0.37	69.43 ± 1.40
DenseNet-121	61.30 ± 0.51	82.70 ± 1.03	63.24 ± 0.20	71.67 ± 0.52
[Tiulpin et al., 2018](1)	79.65 ± 0.66	83.83 ± 1.85	73.36 ± 0.39	78.25 ± 0.89
DRAE	82.53 ± 0.87	88.23 ± 1.52	79.22 ± 0.43	83.48 ± 0.86

- (1) using our samples with the ROIs extracted according to [Tiulpin et al., 2018] in the binary case (KL-G0 vs. KL-G2)

need to be confirmed with further tests. Nonetheless, these results give additional information about the tibial regions most affected by structural changes in the trabecular bone, which is also supported by [Woloszynski et al., 2010, Janvier et al., 2017]. This information can be very useful in the clinical routine to understand early trabecular bone changes due to OA.

4.5 Summary

In this chapter, we have proposed a new representation learning model for knee OA detection from X-ray images. The proposed framework, called Discriminative Regularized Auto-Encoder (DRAE) is based on Auto-Encoders. A discriminative penalty term was introduced in the cost function to maximize the separability between healthy and OA cases. Experimental results were conducted on the multicentric OAI database, which indicates that our method is robust toward artifacts and data acquisition settings. Five ROIs representing the microarchitecture of the lateral, middle, and medial tibial trabecular bone were used for experiments. Obtained results show that the most discriminative regions were located in the medial compartment of the knee, which was also shown in other studies [Woloszynski et al., 2010, Janvier et al., 2017]. This study proves that it is still possible to improve classical Auto-Encoders for better extraction of discriminative patterns linked to OA.

5

Discriminative Shape-Texture Convolutional Neural Networks

Contents

5.1 Introduction and Motivation

Deep Learning approaches based on Convolutional Neural Networks have shown promising results in knee OA detection (see section 2.3.2). Despite this success, the problem of early knee OA detection from plain radiographs remains a challenging task. This is due to the high similarity between OA and non-OA images in the early stage and the CNN architecture nature that neglects the texture information related to the microarchitecture bone changes in their classification layers. More specifically, several studies [Wen et al., 2016, Cai et al., 2018, Cheng et al., 2018] showed that in the case of high inter-class similarities or high intra-class variations, and when using only the softmax cross-entropy loss, features learned with traditional CNNs from the same class are often scattered, and those learned from different classes are overlapped. Moreover, classical CNN models lead to extracting complex correlations in the top layers, corresponding to the overall shape information and neglecting the fine details of the image corresponding to texture information [Cimpoi et al., 2015, Andrearczyk and Whelan, 2016]. Moreover, other studies [Zeiler and Fergus, 2014, Springenberg et al., 2014] have shown that the architecture of classical CNNs tends to increase the abstraction level of the representation with depth. The first layers of the CNNs are designed to learn low-level features, such as edges and curves, which characterize the texture information, while the deeper layers are trained to capture more complex and high-level patterns, such as overall shape information. Consequently, as a result of successive convolutions, activation functions, and pooling operations, the fine image details related to the texture in the top layers of the network disappear. As mentioned above, knee OA is depicted by both shape and texture properties across the overall knee joint. Furthermore, in the early stages of the disease, X-rays are often very similar. It is thus important that the proposed model takes these issues into account. The present work aims to improve the early automatic diagnosis of knee OA using Convolutional Neural Network. First, inspired by previous research in texture CNN [Cimpoi et al., 2015, Andrearczyk and Whelan, 2016] and our recently proposed discriminative regularization [Nasser et al., 2020], we introduce a deep CNN model called DCNN (Discriminative Convolutional Neural Network) [Nasser et al., 2022] to consider both shape and texture changes and maximize separability between OA and non-OA subjects. Next, we propose an extension of the DCNN to a Discriminative Shape-Texture Convolutional Neural Network (DST-CNN) [Nasser et al., 2023] to enhances classification results by (i) improving the quality of texture information before combining it with the overall shape of the knee and (ii) incorporating a new discriminative loss to improve the separability of knee OA classes (KL-0, KL-1, and KL-2) in a multi-class classification way. The potential of our proposed networks was evaluated for both binary and multi-class classification tasks. The main contributions of this study are as follows:

- Development of a new CNN-based networks, able to learn highly discriminative representations and fuse both shape and texture features, to offer relevant classification results.

- The robustness of the proposed networks is evaluated against artifacts and data acquisition settings by conducting experiments on two large public databases: the Multi-center Osteoarthritis Study (MOST) database [Segal et al., 2013] used for training and the OsteoArthritis Initiative (OAI) database [Peterfy et al., 2008] used for validation and tests.

- An ablation study is performed to evaluate the impact of each component of the proposed network on the learning process.

- Transparency in the decision process is ensured thanks to activation maps showing the contribution of the different areas of the knee to the final decision.

This chapter is divided into two main sections: Section 5.2 presents the basic discriminative CNN proposed for the binary classification of knee OA, and Section 5.3 describes the DST-CNN with relevant modifications.

5.2 Discriminative Convolutional Neural Network

This section presents a preliminary study in which we introduce a new deep convolutional neural network called DCNN based on the standard DenseNet model [Huang et al., 2017] to automatically distinguish between the definite absence (KL-0) and the definite presence (KL-2) of knee OA from X-ray images.

5.2.1 Proposed DCNN Network

The proposed network is derived from the classical DenseNet architecture described in A.1.1.F. To tackle the high similarity between OA and non-OA knee X-ray images at the early stages and to better detect the early signs of OA, we force the proposed network to : (i) learn a deep discriminative representation and (ii) take into account the texture-rich features present in the intermediate layers.

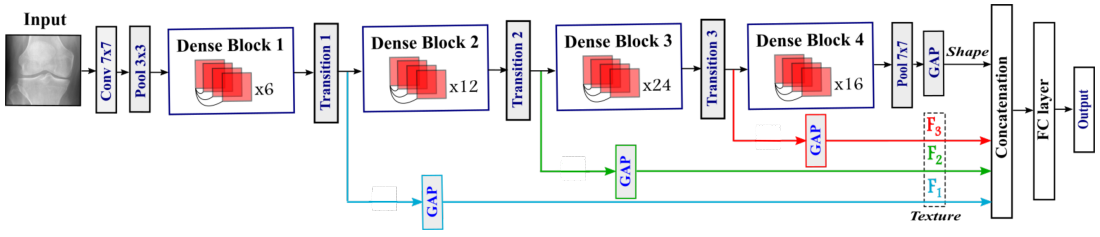


Figure 5.1: Overview of the proposed DCNN network. Combination of mid-level representations and shape information to improve the prediction of OA in early stage. F_l is the global average pooling of the output of the l^{th} transition layer.

5.2.1.A Learning a Deep Discriminative Representation

To learn deep discriminative features, a penalty term is imposed on the mid-level representations of the DenseNet. Apart from minimizing the standard classification loss, the objective is to improve the discriminative power of the network by forcing the representations of the different classes to be mapped faraway from each other. More specifically, we incorporate an additional discriminative term to the original classification cost function. The new objective function, $\mathcal{L}_T$ consists of two terms including the softmax cross-entropy loss and the discriminative penalty one:

$$\mathcal{L}_T = \mathcal{L}_C + \lambda \mathcal{L}_D \tag{5.1}$$

where λ is a trade-off parameter which controls the relative contribution of these two terms.

$\mathcal{L}_C$ is the softmax cross-entropy loss, which is the traditional cost function of the DenseNet model. It aims at minimizing the classification error for each given training sample. Over a batch X of multiple samples of size N, the binary CE loss is defined as:

$$\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{5.2}$$

$\mathcal{L}_D$ represents the discriminative loss used to enforce the discriminative ability of the proposed model. $\mathcal{L}_D$ attempts to bring "similar" inputs close to each other and "dissimilar" inputs apart. To compute $\mathcal{L}_D$, we first feed the set of training samples X to the network and compute the outputs (feature maps) in each layer for each training sample, $x_i \in X$. Then, we compute $F_l(x_i)$, the Global Average Pooling (GAP) of the output feature maps of each transition layer l. Finally, the total discriminative loss $\mathcal{L}_D$ is defined as follows:

$$\mathcal{L}_D = \frac{2}{N}\sum_{l=1}^{L}\sum_{i=1}^{N/2} E_l(F_l(x_i^n), F_l(x_i^p)) \tag{5.3}$$

where E_l is the discriminative loss at a transition layer l. $F_l(x_i^n)$ and $F_l(x_i^p)$ are the GAP of the output feature maps at transition layer l corresponding to the i^{th} negative subject x_i^n (healthy) and positive subject x_i^p (OA), respectively.

In the current study, we evaluate two loss functions, the Triplet loss [Schroff et al., 2015] and the Ω_{disc} loss used in our previous study [Nasser et al., 2020].

The Triplet loss aims to ensure that the image x_i^a (anchor) is closer to all images x_i^p (positive) belonging to the same class, and is as far as possible from the images x_i^n (negative) belonging to an other class. Hence, when using a triple loss, E_L can be defined as:

$$E_l = \sum_{i=1}^{N} \max(d(F_l(x_i^a), F_l(x_i^p)) - d(F_l(x_i^a), F_l(x_i^n)) + \epsilon, 0) \tag{5.4}$$

where d is a distance metric, ϵ is a margin that is enforced between positive and negative pairs.

The Ω_{disc} loss, attempts to encourage classes separability, at each transition layer l, by maximizing the distance between the means μ_l^p and μ_l^n of the learned feature sets ($F_l(x_i^p)$ and $F_l(x_i^n)$) of each class and minimizing their variances v_l^p and v_l^n. The discriminative loss E_l which will be minimized in the use case of Ω_{disc} is defined then as:

$$E_l = \frac{v_l^p + v_l^n}{|\mu_l^p - \mu_l^n|^2} \tag{5.5}$$

5.2.1.B Merging multi-scale feature

As mentioned above, several studies have shown that the first layers of CNNs are designed to learn low-level features, such as edges and curves, which characterize the texture information, while the deeper layers are learned to capture more complex and high-level patterns, such as the overall shape information (see section 3.3.3.A). Moreover, CNN layers are highly related to filter banks methods widely used in texture analysis, with the key advantage that the CNN filters learn directly from the data rather than from handcrafted features (see section 3.3.3.C).

Based on these remarks and especially on the main idea of the Texture CNN (T-CNN) learning model proposed in [Andrearczyk and Whelan, 2016], we propose a simple and efficient modification to the DenseNet architecture to improve its ability to consider both texture and shape information.

Figure 5.1 illustrates the proposed architecture for combining features from the mid-level layers containing texture information with the deep features in the top convolutional layer, containing shape information. By doing so, the network can learn texture information as the overall shape from the input image. This combination of features at different hierarchical layers enables a description of input images at different scales.

5.2.2 Experimental setup

5.2.2.A Data description

Knee X-ray images used to train and evaluate the proposed model were obtained from MOST and OAI studies described in section 2.2. The model was trained using ROIs corresponding to the distal area of the knee extracted from right knees and horizontally flipped left ones. Each ROI was associated with its KL grade. The objective of this study is to distinguish between the definite absence (KL-G0) and the definite presence of OA (KL-G2), which is the most important and challenging task, due to the high degree of similarity between their corresponding X-ray images. KL-G1, is a doubtful grade and was not

Table 5.1: Dataset description and distribution

Group	Dataset	KL-0	KL-2
Train	MOST	6008	3045
Validation	OAI	1116	806
Test	OAI	2313	1545

considered in the current study. Table 5.1 summarizes the number of training, validation and testing samples.

5.2.2.B Implementation details

Our experiments were conducted using Python with the framework Tensorflow on Nvidia GeForce GTX 1050 Ti with 4 GB memory. The proposed approach was evaluated quantitatively using four metrics: Accuracy (Acc); Precision (Pr); Recall (Re) and F1-score (F1).

Dataset preparation

As shown in Table 5.4, data are imbalanced. To overcome this issue during the training stage, data were balanced using the oversampling technique. To do so, different random linear transformations were applied to the samples, including: (i) random rotations using a random angle varying from -15^0 to 15^0, (ii) color jittering with random contrast and brightness with a factor of 0.3, and (iii) a Gamma correction.

Training phase

As mentioned previously, DenseNet pre-trained on ImageNet [Deng et al., 2009] was retained as our basic network structure (see A.1.1.F). The input size of the ROIs is 224×224, which is the standard size used in the literature. The proposed model was trained and optimized end-to-end using Adam optimizer with an initial learning rate of 0.0001. Hyper-parameters (λ, batch size, size of the fully connected layer, ration of dropout) were tuned using grid search on the validation set.

5.2.3 Experimental results

In this section, the performance of our proposed method is evaluated for early knee OA detection. Firstly, two discriminative loss functions are evaluated. Then, the proposed network is compared to some of DL pre-trained models. Finally, a visualisation analysis using t-SNE scatter plots is performed.

We test Triplet Hard and SemiHard losses with three distance metrics: l^2-norm, squared l^2-norm and the cosine similarity distance. We also evaluate the discriminative loss Ω_{disc} proposed in [Nasser et al., 2020]. The results are reported in Table 5.2. As can be seen, the best overall classification performance is obtained using the Ω_{disc} discriminative loss with an accuracy rate of 87.69%. In term of the F1-score,

the highest value (87.06%) is also reached using the Ω_{disc} discriminative loss, which corresponds to a precision rate of 87.48% and recall rate of 86.72%. We notice that Triplet SemiHard loss with l^2-norm distance achieves competitive performance with Ω_{disc} loss. These results show that Ω_{disc} discriminative loss, leads generally to better performance compared to other tested losses. Hence, it is retained for the following experiments.

Table 5.2: Classification Performance of the proposed method using different discriminative loss functions

Discriminative Loss	**Distance Metric**	**Acc (%)**	**Pr (%)**	**Re (%)**	**F1 (%)**
Triplet Hard	l^2-norm	86.21	85.51	86.31	85.82
	squared l^2-norm	86.50	85.94	85.93	85.94
	cosine similarity	86.39	85.76	86.02	85.88
Triplet SemiHard	l^2-norm	87.48	**87.88**	85.94	86.66
	squared l^2-norm	86.91	86.74	85.82	86.21
	cosine similarity	85.82	85.16	85.49	85.31
Ω_{disc}	-	**87.69**	87.48	**86.72**	**87.06**

The proposed method is compared to some DL pre-trained networks, that are the standard DenseNet, ResNet as well as Inception-V3 discribed in Appendix A. Results are reported in Table 5.3. As can be seen, the proposed model achieved the highest prediction performance compared to the other networks. In terms of accuracy, our proposed method obtains a score of 87.69% compared to 85.07%, 86.49% and 84.03% achieved by ResNet-101, DenseNet-169 and Inception-V3, respectively. The highest F1-score (87.06%) is obtained also by our proposed model. Even though DenseNet-169 achieved a high precision compared to other networks, it still has a low recall (75.08%). Therefore, with the exception of the precision values of DenseNet-169, our approach outperforms all other networks for all four metrics. In particular, a significant improvement in terms of F1-score is observed, as our model increases results by 5.14% from the 81.92% achieved by the standard DenseNet to 87.06% for the proposed model.

Table 5.3: Comparison of the proposed method to the pretrained deep learning networks.

Methods		**Acc (%)**	**Pr (%)**	**Re (%)**	**F1 (%)**
ResNet	ResNet-50	83.23	88.41	74.49	80.85
	ResNet-101	85.07	83.56	80.04	81.76
	ResNet-152	84.86	75.99	84.64	80.08
DenseNet	DenseNet-121	85.66	82.76	81.10	81.92
	DenseNet-169	86.49	**89.50**	75.08	81.66
	DenseNet-201	84.76	86.22	73.72	79.48
Inception	Inception-V3	84.03	83.39	75.08	79.02
Proposed model		**87.69**	87.48	**86.72**	**87.06**

In addition to the quantitative evaluation, we check whether our model is able to increase the seg-

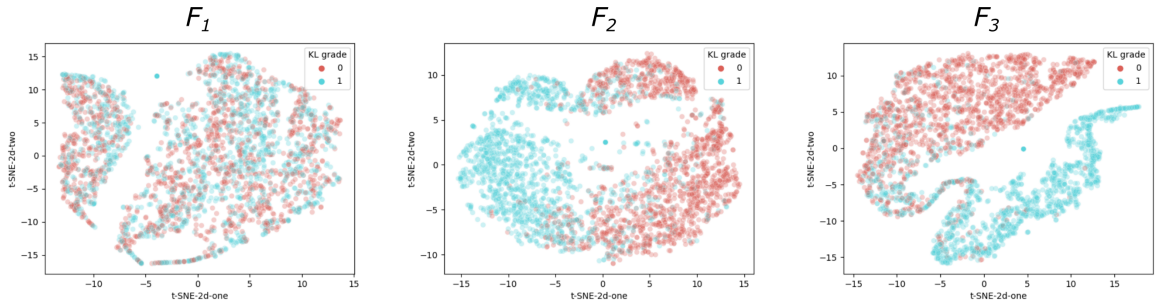


Figure 5.2: Obtained t-SNE scatter plots for each feature levels using our proposed network.

regation of classes. To this end, we display the 2D scatter plots using t-SNE on each features levels $\{F_1, F_2, F_3\}$. Results are illustrated in Figure 5.2. The first column shows the feature vector F_1 extracted from the first transition layer. As can be seen, the two classes significantly overlap. This may be due to common textual features shared between classes, such as edges and contours that form the overall joint shape. The second column shows the learned feature vectors F_2 obtained from the second transition layer. In this case, the network improves the separation between the two classes but not enough. The last column shows the learned features vector F_3 obtained from the third transition layer. Results show that by going deeper, our proposed model learned two discriminant representations. Thus, it leads to a better classes discrimination and thus a good prediction of knee OA at an early stage.

5.3 Discriminative Shape-Texture CNN (DST-CNN)

In this section, we present our proposed Discriminative Shape-Texture CNN (DST-CNN), which improves the DCNN in two ways: (i) it incorporates a new block called GMD that enables the network to extract more powerful texture features (see Figure 5.3), and (ii) it enhances the discriminative ability of the learning model for multi-class classification (see Figure 5.5).

5.3.1 Learning texture representation

To improve texture feature learning, a new block is first incorporated by computing the correlations between feature maps in several mid-level layers. More specifically, the feature maps corresponding to an input training image $x \in X$ are computed at each layer l of the network. A layer l with K_l filters has a set $\{f_k^l | k = 1, ..., K_l\}$ of feature maps each of which has $M_l = H_l \times Wl$ elements. Then, each feature map f_k^l is flattened and stored in a matrix $F^l \in \mathbb{R}^{K_l \times M_l}$, where $F_{k,m}^l$ is the m^{th} element of the k^{th} feature map in layer l (see Figure (5.4, step (A))). To capture texture properties, some modules of our DST-CNN network need to be agnostic to spatial information. Thus, contrary to what is proposed in [Andrearczyk and Whelan, 2016], we suggest computing correlations between the different extracted features before feeding them to the classification layers. To this end, our solution consists in introducing a Gram Matrix

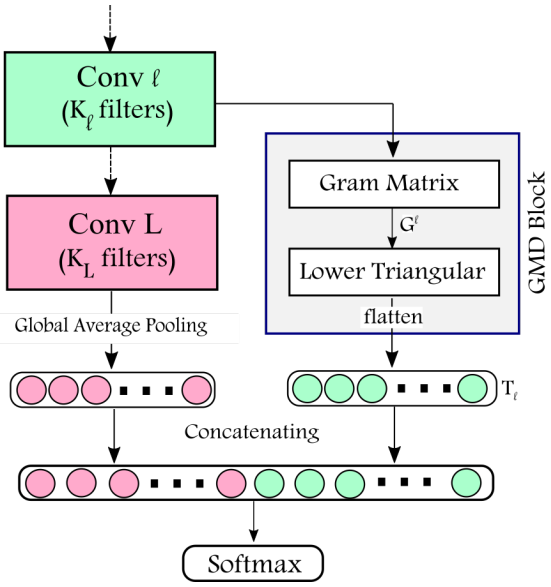


Figure 5.3: A visualization of the texture learning process using the proposed GMD block. At layer l, the input of the GMD block is a matrix F_l of K_l feature maps with spatial dimensions of $K_l \times M_l$. The output T_l is obtained by flattening the lower triangular part of the computed Gram matrix G_l. The resulting texture representation T_l is concatenated with the global average pooling of the last convolutional layer and passed to the softmax layer to obtain class predictions.

Descriptor (GMD) block that takes F^l as input and computes feature correlations $G^l \in \mathbb{R}^{K_l \times K_l}$ using the Gram matrix [Gatys et al., 2015] (see Figure (5.4, step (B))), where G^l_{ij} is the inner product between the flattened feature maps i and j in the layer l, computed as follows:

$$G^l_{ij} = \sum_{m}^{M_l} F^l_{i,m} F^l_{j,m} \tag{5.6}$$

However, when extracting the Gram matrix representations from various layers of the DST-CNN network, the number of parameters of the model becomes very large. As the Gram matrix is symmetric, we selected only the lower triangular part of each G^l matrix (see Figure (5.4, step (C))). Consequently, the number of parameters was reduced from $K_l \times K_l$ to $K_l \times (K_l + 1)/2$.

Once the lower triangular part of G^l has been selected, we flatten these elements into a single vector named T_l. Next, we use a concatenation layer to fuse the set of texture descriptors $\{T_l | l = 1, .., L\}$ from the GMD blocks with the global average pooling of the last convolutional layer that represents the overall shape information. Finally, this vector is fed to the final classification layer (See Figure 5.3 and 5.7).

5.3.2 Learning discriminative representation

Once the texture and shape information have been considered, we focus on training the model to distinguish between knee OA grades: KL-0, KL-1, and KL-2 (Figure 5.8).

To this end, we introduce a discriminative penalty term on the mid-level texture representations

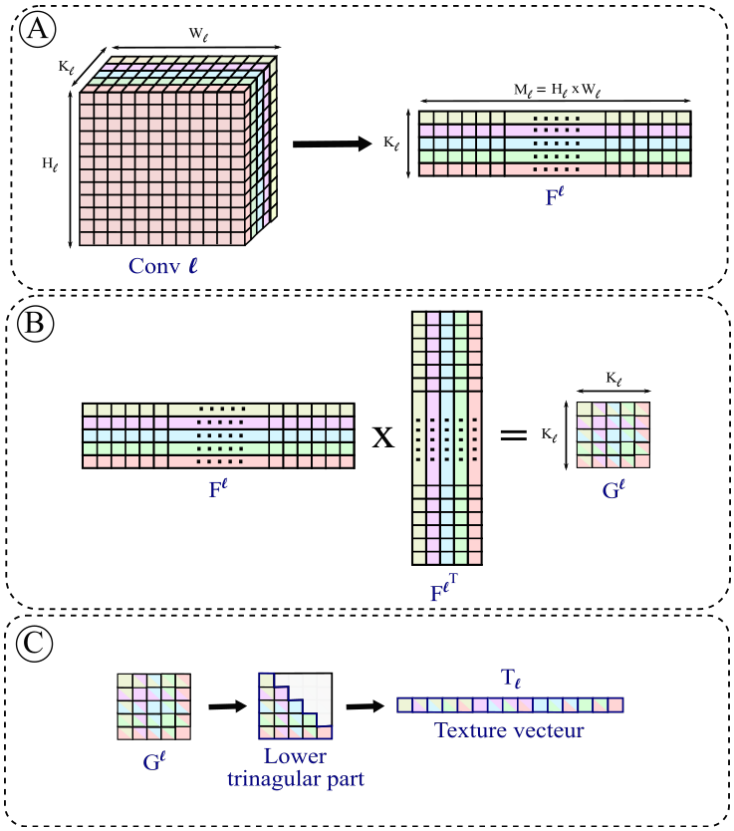


Figure 5.4: Visualized GMD Block steps to calculate the texture representation T_l from the feature maps of a convolutional layer l. **Step (A)**: flattened and stored the set of feature maps in a matrix F^l, **Step (B)**: computed the Gram matrix representations G^l by multiplying F^l with its transpose, **Step (C)**: selected and flattened the lower triangular part of G^l to construct the resulting texture representation T_l.

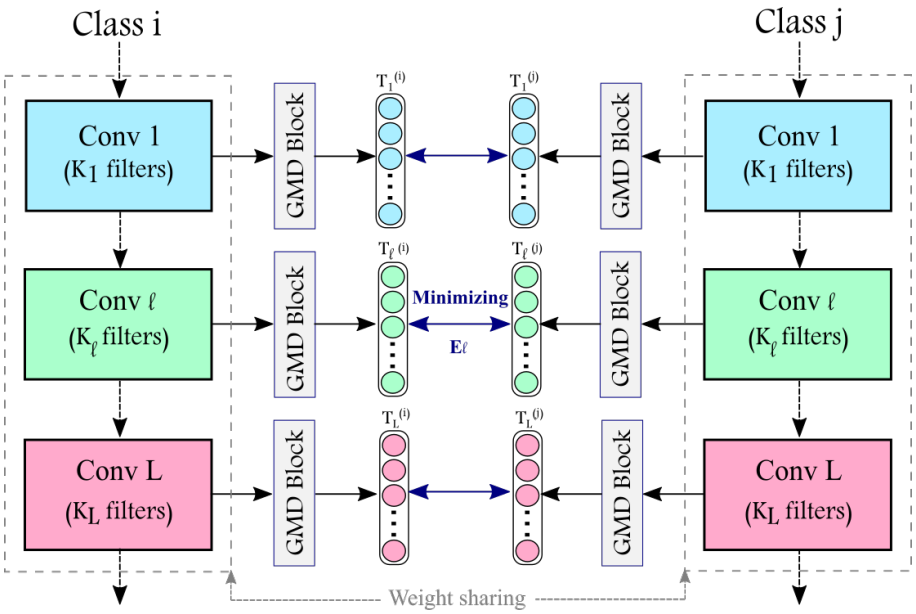


Figure 5.5: A visualization of the discriminative learning process of the proposed DST-CNN network. The input images of class i and j are passed through the network and their corresponding representations $T_l^{(i)}$ and $T_l^{(j)}$ are computed via the GMD blocks. The discriminative loss E_l is computed at each layer l to maximize the separability between feature classes.

(see. Figure 5.5). Beyond minimizing the standard classification loss, the objective is to improve the discriminative power of the network by forcing the computed texture descriptors $\{T_l | l = 1, .., L\}$ of the different classes to be mapped far away from each other. More specifically, we incorporate an additional discriminative term in the original classification cost function. The new objective function J_T consists of two terms including the softmax Cross-Entropy (CE) loss and the discriminative penalty one:

$$J_T = J_{Cls} + \lambda J_{Disc} \tag{5.7}$$

where J_{Cls} is the softmax CE loss (Eq. (5.9)), J_{Disc} represents the new discriminative loss, and λ is a trade-off parameter that controls the relative contribution of the two terms.

J_{Cls} loss aims at minimizing the classification error for each given training sample. It evaluates how well the network prediction corresponds to the target classification. Over a batch X of multiple samples of size N with C classes, the CE loss is defined as:

$$J_{Cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} \log(\hat{y}_{i,j}) \tag{5.8}$$

where $y_{i,j} \in \{0, 1\}$ is the target class label which indicates if j is the correct label of the i^{th} sample; and $\hat{y}_{i,j} \in [0, 1]$ is the output probability that the sample i belongs to class j.

J_{Disc} is used to enforce the discriminative ability by forcing the DST-CNN model to learn a feature

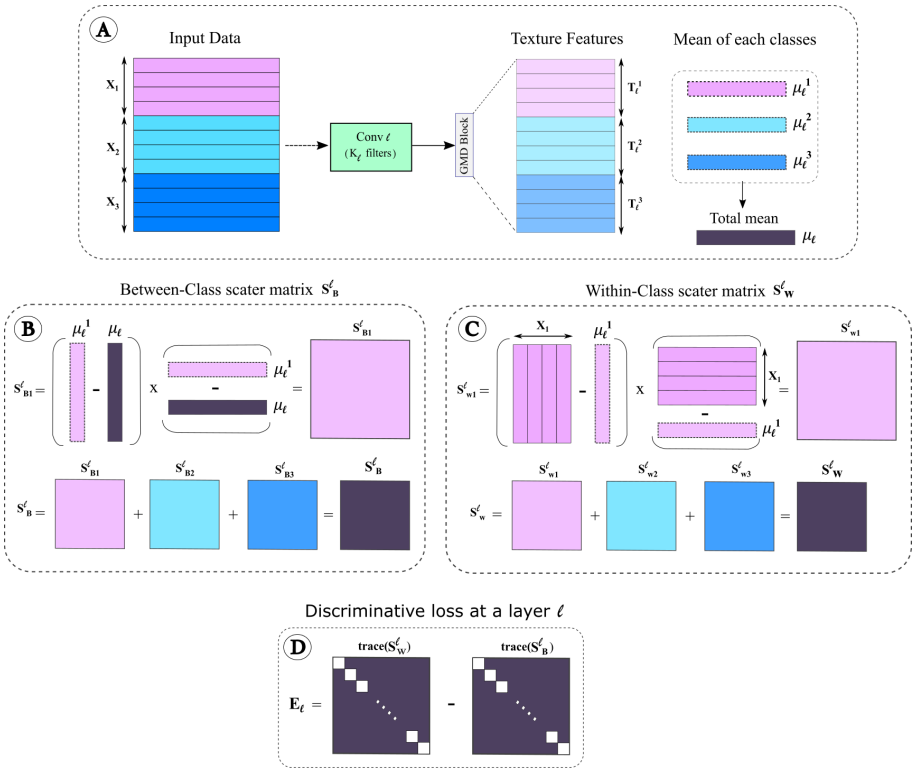


Figure 5.6: Visualized steps to calculate the discriminative loss E_l at a layer l.

space in which similar examples are close while dissimilar ones are far apart. For example, images x_i belonging to the same class c are pulled together, while images from distinct classes are pushed apart from each other. In other words, J_{Disc} attempts to minimize the within-class scatter, denoted by S_w, while it maximizes the extra-class scatter, denoted by S_b (see Figure (5.6, steps (B & C))). S_w^l and S_b^l at the l^{th} layer are computed as:

$$S_w^l = \sum_{c}^{C} \sum_{x_i \in X_c} (T_l(x_i) - \mu_l^{(c)})(T_l(x_i) - \mu_l^{(c)})^t \tag{5.9}$$

$$S_b^l = \sum_{c}^{C} n_c(\mu_l^{(c)} - \mu_l)(\mu_l^{(c)} - \mu_l)^t \tag{5.10}$$

where,

$$\mu_l^{(c)} = \frac{1}{n_c} \sum_{x_i \in X_c} T_l(x_i) \qquad \mu_l = \frac{1}{N} \sum_{x_i \in X} T_l(x_i) \tag{5.11}$$

n_c is the number of samples of the class c. $\mu_l^{(c)}$ and μ_l represent the mean vector of the extracted features T_l from a layer l over a batch of size N corresponding to the c^{th} class and of all classes, respectively (see Figure (5.6, step (A))).

The contribution of the l^{th} layer to the total discriminative loss is then given by:

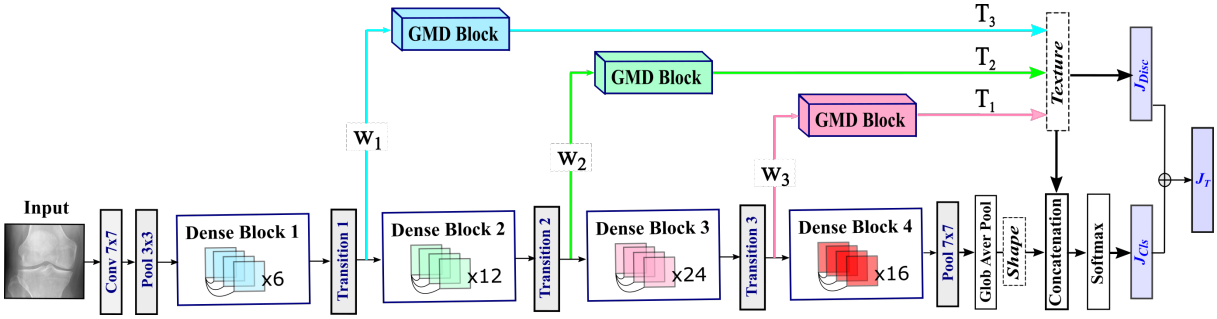


Figure 5.7: Overview of the proposed DST-DNet network combining texture and shape information to improve the early detection of OA. T_l represents the texture feature of the l_{th} transition layer, w_l is a binary factor controlling the contribution of the features at the l_{th} transition layer.

$$E_l = tr(S_w^l) - tr(S_b^l) \tag{5.12}$$

where $tr(.)$ is the matrix trace operator as shown in Figure (5.6, steps (D)). By minimizing E_l, our network ensures that features of the same class are close, and those of different classes are distant.

Finally, the total discriminative loss J_{Disc} is given by:

$$J_{Disc} = \sum_{l=1}^{L} w_l E_l \tag{5.13}$$

where $w_l \in \{0, 1\}$ is the weight factor corresponding to the contribution of layer l to the total discriminative loss.

5.3.3 Architecture and training process

In this study, we used DenseNet-121 as our network backbone. So, our proposed variant of the model is referred to as Discriminative Shape-Texture DenseNet (DST-DNet). Figure5.7 provides a schematic overview of the architecture of the DST-DNet. While training the classical DenseNet network requires only optimizing the classification objective function, J_{Cls}, training the DST-DNet network requires the optimization of the additional discriminative term J_{Disc} (Eq. 5.13). Gradients of J_{Disc} and thus the gradient of J_T, ∇J_T, can be computed using the standard error back-propagation. Once ∇J_T has been computed, the set W of DST-DNet parameters is updated according to the following equation:

$$W^{t+1} = W^t + \alpha \nabla J_T \tag{5.14}$$

Figure 5.8: Different X-ray images of knees with different KL grades showing the high similarity between grades KL-0, KL-1, and KL-2.

5.4 Experimental settings

5.4.1 Dataset

Knee X-ray images used to train and evaluate the proposed approach were obtained from two public databases: The Multicenter Osteoarthritis Study (MOST) and the OsteoArthritis Initiative (OAI) (see section 2.2). The entire MOST database (3026 subjects) was used for training, and the OAI baseline database (4796 subjects) was used for validation and testing. The model was trained with Regions Of Interest (ROI) corresponding to the distal area of the knee extracted from right knees and horizontally flipped left ones (see Figure 5.8). Each ROI was associated with its KL grade. Table 5.4 summarizes the number of training, validation, and testing samples.

Table 5.4: Dataset distribution

Group	Dataset	KL-0	KL-1	KL-2
Train	MOST	6008	2933	3045
Validation	OAI	1116	513	806
Test	OAI	2313	1071	1545

5.4.2 Evaluation metrics

In our experiments, quantitative evaluation was performed using five different metrics:

- Accuracy (Acc): is the percentage of predictions that match exactly the ground-truth (Eq. 4.5);
- Precision (Pr): is the fraction of true positives among predicted positives (Eq. 4.6);
- Recall (Re): is the fraction of a total number of true positives retrieved (Eq. 4.7);
- F1-score (F1): also called F-measure, is the harmonic mean of precision and recall (Eq. 4.8);

- AUC (computed Area Under the Curve): represents the degree or measure of separability, it is computes the classification performance of the model across different thresholds and provides an aggregate measure. AUC ranges in value from 0 to 1. In general, an AUC of 0.5 suggests no discrimination, 0.6 to 0.7 is considered acceptable, 0.7 to 0.8 is considered good, 0.8 to 0.9 is considered very good, and more than 0.9 is considered excellent.

5.4.3 Implementation details

Our experiments were performed using the TensorFlow framework and runs on a server with a single Nvidia GPU card (TESLA A100, 40 GB memory).

Dataset preparation: As shown in Table 5.4, the data are imbalanced. To overcome this issue during the training stage, the data were balanced using different random linear transformations such as random rotations (-15^0 to 15^0); color jittering with random contrast and brightness (factor of 0.3); and Gamma correction.

Training phase: All images were resized to 224×224 before being fed to the proposed network, which is the standard input size of the backbone network. The DST-DNet model was trained and optimized end-to-end using Adam optimizer with an initial learning rate of 0.0001. Hyperparameters (λ and batch size) were tuned using a grid search on the validation set. We retained the values for which the models performed best in terms of F1-score and AUC on the validation set.

5.5 Experimental results

In this section, the performance of the proposed DST-DNet is evaluated for early knee OA prediction. The DST-DNet model is compared to SoA CNN DL models (see section A.1.1), including the standard DenseNet-121, ResNet-50, Xception, EfficientNet as well as to MobileNet, in both binary and multi-class classification tasks. In addition, our model is compared to three SoA knee OA diagnosis CNN-based models proposed by Tiulpin *et al.* [Tiulpin et al., 2018], Chen *et al.* [Chen et al., 2019], and Nguyen *et al.* [Nguyen et al., 2020] in the multi-class classification task.

5.5.1 Binary classification results

First, the models were trained to detect two classes. Therefore, three classification scenarios were performed: Normal Patients (KL-0) vs. Mild OA cases (KL-2), Normal Patients (KL-0) vs. Doubtful OA cases (KL-1), and Doubtful OA cases (KL-1) vs. Mild OA cases (KL-2). Tables 5.5, 5.6, and 5.7 summarizes the results of the binary classification experiments.

Table 5.5: KL-0 vs KL-2 classification performance (%).

Methods	Acc	Pr	Re	F1	AUC
DenseNet-121	86.37	82.27	84.07	83.16	92.70
ResNet-50	85.69	81.40	83.30	82.34	92.59
Xception	85.15	88.03	72.81	79.70	92.33
EfficientNet	85.43	**89.41**	72.16	79.87	90.72
MobileNet	84.89	84.80	75.86	80.08	91.97
DST-DNet	**87.92**	88.04	**86.66**	**87.21**	**94.16**

Table 5.6: KL-0 vs KL-1 classification performance (%).

Methods	Acc	Pr	Re	F1	AUC
DenseNet-121	72.22	55.36	63.21	59.02	75.80
ResNet-50	72.73	57.86	50.89	54.15	73.01
Xception	72.63	56.91	55.74	56.32	74.53
EfficientNet	70.12	52.20	66.39	58.45	73.56
MobileNet	71.28	54.15	60.32	57.07	75.12
DST-DNet	**74.08**	**68.46**	**69.85**	**69.01**	**76.38**

5.5.1.A KL-0 vs. KL-2

A comparison of the evaluated DL networks shows that EfficientNet has better precision (89.41%) than DenseNet-121 while DenseNet has better accuracy (86.37%), recall (83.16%), F1-score (83.16%), and AUC (92.70%). Except for precision, DST-DNet outperforms all the other networks for all selected metrics. In particular, compared to the standard DenseNet-121, a remarkable improvement is noticed in terms of recall and F1-score, as our model increases the results by 2.59% and 4.05%, respectively.

5.5.1.B KL-0 vs. KL-1

Classification of KL-0 vs. KL-1 is the most difficult task. Table 5.6 shows that DenseNet outperforms the other SoA DL networks with F1-score (59.02%), and AUC (75.80%). EfficientNet reaches the highest recall (66.39%) but also the lowest accuracy (70.12%) and precision (52.20%). Concerning our network, DST-DNet achieves the highest classification performance in terms of accuracy (74.08%), precision (68.46%), recall (69.85%), F1-score (69.01%), and AUC (76.38%). DST-DNet also provides the most balanced results, with a significant improvement for all five metrics.

5.5.1.C KL-1 vs. KL-2

As shown for (KL-0 vs. KL-1) and (KL-0 vs. KL-2), our proposed network provides the most balanced classification results (see Table 5.7). DST-DNet achieves the highest accuracy (75.04%), recall (75.05%), F1-score (74.84%), and AUC (82.65%), while ResNet-50 achieves the highest precision (85.15%) but the lowest recall (57.54%) and F1-score (68.68%). The lowest accuracy (68.27%) and AUC (76.40%) are

Table 5.7: KL-1 vs KL-2 classification performance (%).

Methods	Acc	Pr	Re	F1	AUC
DenseNet-121	71.10	82.31	65.05	72.67	79.45
ResNet-50	69.00	**85.15**	57.54	68.68	78.82
Xception	69.64	83.32	60.78	70.28	79.49
EfficientNet	68.27	81.06	60.39	69.21	76.40
MobileNet	69.57	81.03	63.30	71.08	78.02
DST-DNet	**75.04**	75.90	**75.05**	**74.84**	**82.65**

Table 5.8: Multi-class classification (KL-0, KL-1, KL-2) performance (%)

Methods	Acc	Pr	Re	F1	AUC
DenseNet-121	65.15	64.10	60.53	61.36	79.78
ResNet-50	62.02	**66.83**	58.90	58.87	80.33
Xception	62.61	64.97	57.13	57.71	79.70
EfficientNet	62.02	65.33	59.20	59.56	80.39
MobileNet	62.81	63.35	59.39	59.99	78.48
DST-DNet	**68.70**	65.40	**62.92**	**63.67**	**82.97**

obtained by EfficientNet.

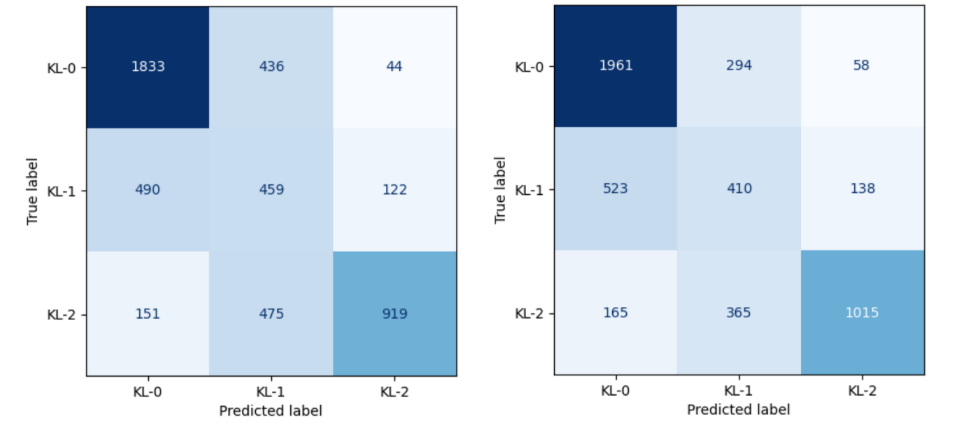
5.5.2 Multi-class classification Results

5.5.2.A Comparison with DL SoA networks

After testing the models in the binary case, we trained and evaluated these models in the case of a multi-class classification task. Table 5.8 summarizes the results obtained when grouping the three classes (KL-0, KL-1, and KL-2). DST-DNet obtained a score of 68.70%, 63.67%, and 82.96% in comparison to DenseNet, which reached 65.14%, 61.36%, and 79.78% in terms of accuracy, F1-score, and AUC, respectively. The lowest accuracy (60.97%), F1-score (57.69%), and AUC (76.19%) were obtained by EfficientNet.

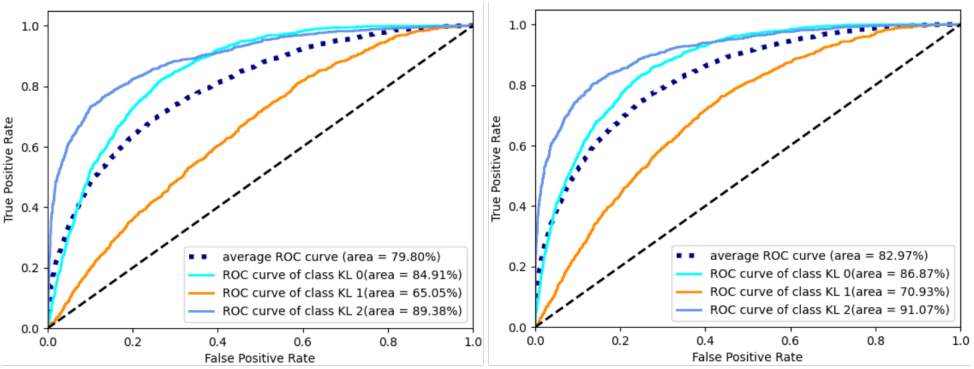
For more comparisons in the case of the multi-class classification, Figure 5.9 shows the confusion matrices for the standard DenseNet (Figure 5.9.a) and DST-DNet (Figure 5.9.b). The rate of correct classification of KL-0, KL-1, and KL-2 samples achieved by DenseNet was 79.25%, 42, 86%, and 59, 48%, while for DST-DNet it was 84.78%, 38.28%, and 65.70%, respectively. Therefore, DenseNet detects patients with Doubtful OA (KL-1) slightly better, but it is not as effective as the poposed DST-DNet model which distinguishes between definite absence (KL-0) and definite presence of OA (KL-2).

Figure 5.10 presents the ROC curves for DenseNet (Figure 5.10.a) and for DST-DNet (Figure 5.10.b). As can be seen, the ROC curves resulting from the proposed DST-DNet network are more significant for all classes in comparison to DenseNet. The average AUC values of KL-0, KL-1, and KL-2 classes for DenseNet-121 are 84.91%, 65.05%, and 89.38%, while for DST-DNet, values are 86.87%, 70.93%, 91.07%,



(a) DenseNet-121 (b) DST-DNet

Figure 5.9: Obtained confusion matrices for the multi-class classification task. (a) DenseNet-121 and (b) DST-DNet.



(a) DenseNet-121 (b) DST-DNet

Figure 5.10: Obtained ROC curves for the multi-class classification task. (a) DenseNet-121 and (b) DST-DNet.

respectively.

5.5.2.B Comparison with SoA knee OA diagnostic models

Table 5.12 shows comparisons with three SoA CNN-based models proposed by Tiulpin *et al.* [Tiulpin et al., 2018], Chen *et al.* [Chen et al., 2019], and Nguyen *et al.* [Nguyen et al., 2020], for classifying knee OA using X-ray images. The proposed DST-DNet achieves the highest multi-class classification performance in terms of precision, recall, and F1-score. The model of Chen *et al.* achieved 0.47% higher accuracy than our model but at the same time, it is 4.76% and 4.39% lower in terms of precision and F1-score, respectively. The lowest precision and recall were achieved by [Nguyen et al., 2020] and [Tiulpin et al., 2018], respectively.

These results point out the advantage of improving texture learning and using a discriminative regularization in CNNs for early knee OA diagnosis. In addition, we note that the three models [Tiulpin et al., 2018, Chen et al., 2019, Nguyen et al., 2020] achieve good results when moderate (KL-3) and Severe

Table 5.9: Comparison to recent models on the multi-class classification task (KL-0, KL-1, KL-2)

Models	Acc (%)	Pr (%)	Re (%)	F1 (%)
[Tiulpin et al., 2018]	59.67	62.24	59.67	59.60
[Chen et al., 2019]	**69.17**	60.64	60.81	59.28
[Nguyen et al., 2020]	56.36	61.18	61.87	61.00
Ours	68.70	**65.40**	**62.92**	**63.67**

Table 5.10: Ablative Analysis: Classification performance (%)

J_{Cls}	J_{Disc}	Feature levels			Multi-classification		
		w_1	w_2	w_3	Acc	F1	AUC
✓					65.15	61.36	79.78
✓		✓			65.48	60.43	81.07
✓			✓		66.56	62.63	81.39
✓				✓	66.40	62.92	80.95
✓		✓	✓		65.79	62.43	81.35
✓		✓		✓	66.68	61.18	81.35
✓			✓	✓	66.52	61.43	82.37
✓		✓	✓	✓	66.91	62.74	81.89
✓	✓	✓			67.23	62.30	81.37
✓	✓		✓		67.74	62.18	81.30
✓	✓			✓	67.66	62.03	81.91
✓	✓	✓	✓		67.39	63.28	81.78
✓	✓	✓		✓	68.29	62.10	82.07
✓	✓		✓	✓	68.39	62.59	82.19
✓	✓	✓	✓	✓	**68.70**	**63.67**	**82.97**

(KL-4) OA cases are considered. For example, the performance of the models proposed in [Tiulpin et al., 2018], [Chen et al., 2019], and [Nguyen et al., 2020] decrease, respectively from an accuracy score of 66.71%, 69.60% and 71% (when considering all five KL grades) to 59.67%, 69.17% and 56.36% (in the case of early detection, KL-0, KL-1 and KL-2). This discrepancy in the results explains why we mainly focused on the challenging task of the early diagnosis of knee OA, rather than considering all KL grades in the multi-classification task.

5.6 Analysis

Without losing any generality, in this section, we focus on the multi-class classification task to conduct more experiments and show the potential of the proposed DST-DNet.

5.6.1 Ablative study

In this section, we conduct an ablation study to demonstrate the effectiveness of the different components of the DST-DNet network.

Table 5.11: Classification performance (%) of J_{Disc} vs. other popular losses

Discriminative loss	Acc	Pr	Re	F1	AUC
Triplet loss [Schroff et al., 2015]	67.17	63.21	**62.97**	63.05	82.09
N-pair loss [Sohn, 2016]	66.99	**65.84**	62.51	63.21	81.62
Our J_{Disc} loss	**68.70**	65.40	62.92	**63.67**	**82.97**

5.6.1.A Contribution of each component

To provide more insight into the proposed network, we performed an ablation study to analyze the effect of every component on the final performance. We first investigated the usefulness of the three feature levels (w_1, w_2, w_3) related to texture information by merging them with the shape information in the top classification layers without using the discriminative loss. Then, we evaluated the performance of the model by applying both proposed solutions: (i) merging texture with shape and (ii) adding the discriminative loss to the objective function. Table 5.10 shows the results obtained by the ablative study. As can be seen, a significant improvement is noticed when we gradually add texture representations at different levels without using J_{Disc} loss. The accuracy, F1-score and AUC are increased from 65.15%, 61.36% and 79.78% to 66.91%, 62.74% and 81.89%, respectively. The performance of the model continues to increase when the discriminative loss is associated to the model. The F1-score increases from 60.43%, when $w_1 = 1$ and considering only J_{Cls}, to 62.30%, when adding J_{Disc}. The best performance was reached when considering the three levels of texture features (w_1, w_2, w_3) combined with the softmax CE and the discriminative loss. These experiments show that each component of the DST-DNet (discriminative loss and texture representation at each level, w_l) contributes to the efficiency of the proposed model.

5.6.1.B J_{Disc} vs. other popular losses

We also investigated the performance of the proposed discriminative loss J_{Disc} (Eq. 5.13) compared to the MOST popular losses: Triplet loss [Schroff et al., 2015] and the N-pair loss [Sohn, 2016]. Table 5.11 shows that J_{Disc} loss provides the best performance compared to the other losses. When using the J_{Disc} loss, the DST-DNet reaches an accuracy of 68.70% compared to 67.17% and 66.99% achieved using the Triplet and the N-pair losses, respectively. In terms of F1 and AUC scores, the highest values were also obtained using our DST-DNet along with J_{Disc}. However, the use of the DST-DNet architecture with the Triplet and the N-pair losses still improves the results compared to the standard DenseNet. These results show that in general, the J_{Disc} loss yields better performance compared to the other losses.

5.6.2 Attention maps

In this section, we ensure transparency in the decision process by providing attention maps showing which areas of interest contributed to the model's decision. Gradient-weighted Class Activation Mapping

(Grad-CAM) [Selvaraju et al., 2017] was used to highlight the pixels activated by the neurons through back-propagation. Figure 5.9 shows Grad-CAM maps for DenseNet as well as for DST-DNet. As can be seen, the pixels highlighted by DenseNet are mainly located on the left side of the knee joint. In comparison, our model does not react only to the medial compartment, as most highlighted areas are located on both the left and right sides of the knee joint. This demonstrates that our model learns more local features corresponding to relevant radiological findings including shape (osteophytes and joint space narrowing) as well as texture details.

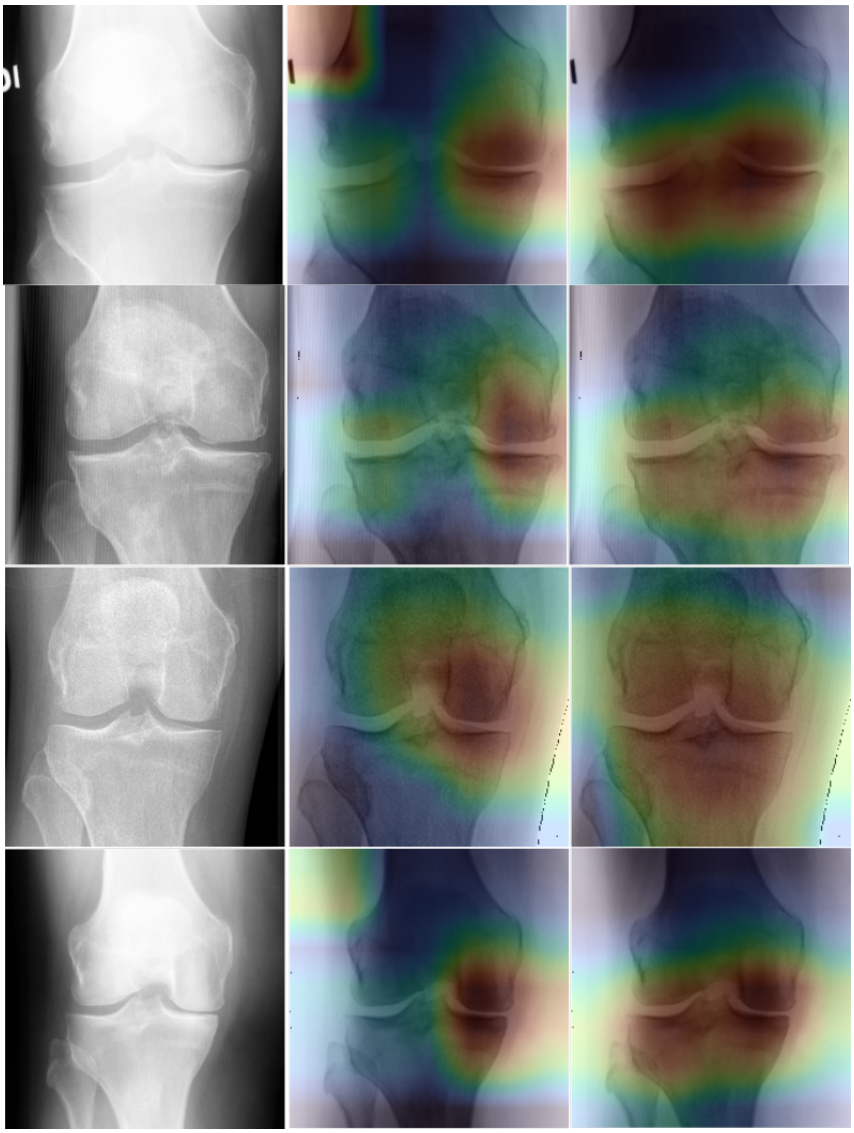


Figure 5.11: Examples of Grad-CAM activation maps obtained for KL-2 patients. X-ray images (a), CAMs: DenseNet (b) and DST-DNet (c).

5.6.3 Discussion

CNN models have shown promising results in several medical imaging applications. Despite this success, the problem of early knee OA diagnosis using CNNs remains a challenging task. This is due to the high similarity between KL OA grades in the early stage and the CNN architecture nature that neglects the texture information related to the bone microarchitecture changes in their classification layers. In this chapter, we introduced a Discriminative Shape-Texture CNN-based model to effectively diagnose knee OA at the early stage.

The main contributions of this work lie in improving the CNN model by incorporating a discriminative loss to improve class separability and to deal with the problem of high inter-class similarities. Furthermore, we enhanced the network texture analysis by introducing a new GMD block to compute the texture features from a multi-scale representation and then combine them with the shape features in the top layers. Experiments on two large public datasets one for training (MOST dataset) and the other for validation and testing (OAI dataset) demonstrate the robustness of the proposed approach.

Compared to the SoA DL models, except for the precision values obtained in (KL-0 vs. KL-2), (KL-1 vs. KL-2), and in multi-class classification experiments, our model outperforms all the evaluated DL models for all metrics (accuracy, precision, recall, F1-score, AUC) (Tables (5.5,5.6,5.7) and Table 5.8). Furthermore, the F1 scores show that the proposed model is robust and provides well-balanced classification scores. Considering that the identification of healthy patients (KL-0) among doubtful ones (KL-1) is a challenging task, our network is still competitive even if a big gap between accuracy and F1-score performance is recorded. This is mainly due to the high similarities between their X-ray images, which may lead to the extraction of the same features. Another possible reason is the imbalance of the number of samples in these two classes, as the number of healthy patients is significantly higher than the number of doubtful cases (Table 5.4). Additionally, this may be related to the confidence granted to the ground truth label associated to doubtful images. As it is doubtful, the label assigned by the network may be more credible. To confirm the obtained results, we showed in Figure 5.9 and 5.10 the confusion matrix and the ROC curve. The obtained AUC scores for each class demonstrate the superiority of our models and confirm its ability to distinguish between all classes.

Compared with three CNN-based SoA knee diagnosis approaches, our model achieves the highest multi-class classification performance in terms of precision, recall, and F1-score. The model of Chen *et al.* achieved 0.47% higher accuracy than our model but at the same time 4.39% lower in terms of F1-score. We note that all three models give good results when Moderate (KL-3) and Severe (KL-4) OA cases are included. For example, the model of Tiulpin *et al.*, Chen *et al.*, and Nguyen *et al.* goes from an accuracy of 59.67%, 69.17%, 56.36% to 66.71%, 69.60%, 71%, respectively. This discrepancy in the results explains why we focus on the challenging task of early knee OA diagnosis rather than the muti-classification task including all KL-grades.

Table 5.12: Comparative performance between our model and recent models on the multi-class classification task (KL-0, KL-1, KL-2)

Models	Acc (%)	Pr (%)	Re (%)	F1 (%)
Tiulpin *et al.* [Tiulpin et al., 2018]	59.67	62.24	59.67	59.60
Chen *et al.* [Chen et al., 2019]	**69.17**	60.64	60.81	59.28
Nguyen *et al.* [Nguyen et al., 2020]	56.36	61.18	61.87	61.00
Proposed DST-DNet	68.70	**65.40**	**62.92**	**63.67**

The results of the ablation study (Section 5.6.1) demonstrated the effectiveness of each component of the proposed DST-DNet model. As noted in Table 5.10, the accuracy increased from 65.48% when using only the first feature level to 66.91% when considering all three feature levels, and the accuracy continues to increase when the discriminant loss is added to the objective function, reaching 68.70%. We have also provided a visualization analysis using the Grad-CAM technique to ensure that the proposed model captures the most relevant features to knee OA diagnosis.

5.7 Summary

In this chapter, we have introduced a new Discriminative Shape-Texture Convolutional Neural Network for the early diagnosis of knee OA using X-ray images. Classic CNNs are unable to take both texture and shape information into account and address the high similarity issue between early knee OA cases. The discriminative ability of our proposed model is improved by incorporating a penalty term into the standard classification objective function of classical CNN models. Both texture and shape information were considered by computing the correlations between feature maps in several intermediate layers and combining them with the global average pooling in the top layers of the network. Our experimental analysis showed an iprovement of the classification performance when using the proposed network, in comparison to SoA networks. Through the Grad-CAM visualization technique, we demonstrated that in most cases, the model learns interesting features relevant to the decision-making process. Additionally, we conducted an ablation study, which showed the usefulness and effectiveness of each component of the proposed network. To the best of our knowledge, this is the first work based on a deep neural network that combines shape and texture features for early automatic diagnosis of knee OA. The proposed approach can be easily integrated into various CNN models and can be used in different medical imaging tasks along with other types of data.

6

Conclusion

Contents

TODAY, early diagnosis of knee osteoarthritis is critical to suggesting effective treatment before facing severe and irreversible pathology and to support and enable patients to address lifestyle factors that influence the disease. In the past decade, deep learning has gained a lot of attention from the research community and has had great success in many medical imaging applications. However, despite these successes, early diagnosis of knee OA from plain radiographs has remained a very challenging task. The need to improve deep learning models to better predicted early symptoms of OA was one of the main motivations that led to the initiation of this thesis. This chapter summarizes the thesis, highlights proposed contributions and discusses several research directions for future work.

6.1 Summary of Contributions

In this thesis, we worked on improving the class separability between Normal (KL-0), Doubtful (KL-1), and Mild OA (KL-2) cases in binary and multi-class classification. For this purpose, the problem of using deep learning to automatically predict the early symptoms and bone changes was studied and approached. The material in Chapter 4 proposed a new representation learning model, called Discriminative Regularized Auto-Encoder (DRAE), for automatically discovering interesting latent representations through classical unsupervised learning and a new discriminative regularization. The discriminative regularization work by forcing the network to capture not only the important features present in the data but also the most discriminative one that minimizes the intra-class variation and maximizes the inter-class distance. This first contribution is a step in the direction of making DL models useful in the case of strong similarity between classes, which we continue to explore in the next Chapters.

Since DRAE is a fully connected network, it fails to capture the patterns in the pixel data because they do not hold the neighborhood information. Moreover, the number of weights increases rapidly with the size of the image becoming unmanageable. While DRAEs focus on a small region below the tibial plateau, OA can occur in any area of the knee joint. In Chapter 5, we described a discriminative convolutional neural network to better capture discriminative patterns of the entire knee joint area. The proposed method improves the quality of early prediction of OA by incorporating discriminative loss into the standard CNN objective function to enforce the separation between the features of OA and non-OA patients, and a multi-scale feature concatenation strategy to improve the representation of the fine properties in the top classification layer.

In the same Chapter, we also address the problem of enhancing texture analysis using the main idea of the discriminative CNN. We introduce a Discriminative Shape-Texture Convolutional Neural Networks (DST-CNN) that extends the previous approach by adding a new block to extract texture information by computing the correlations between feature maps in several intermediate layers. We also adapt the discriminative loss to fit with the multi-class classification task. Using the DST-CNN, we can better characterize the shape and texture information necessary for detecting early OA symptoms and bone

changes. We show that this approach can be easily integrated into various state-of-the-art CNN models and achieves highly competitive results compared to other existing deep learning methods.

6.2 Future Work

Future work that can be addressed to extend this research and further develop the field of discriminative feature learning and texture analysis using deep learning are as follows:

Improving the Discriminative Regularized Auto-Encoder (DRAE) model

Extending the model architecture: The fully connected nature of DRAE presented in Chapter 4 limits its ability to capture patterns in the pixel data that may be influenced by neighborhood information. Future research could investigate the incorporation of convolutional layers or other techniques to exploit the spatial relationships within the knee joint area. This would help the model better understand the local context and potentially improve its performance in detecting OA-related features.

Hybrid reconstruction loss: Another possible direction for enhancing the DRAE model is to incorporate a hybrid reconstruction loss that combines edge reconstruction capabilities with texture preservation. This enhancement can enable the auto-encoder model to capture both structural and textural information, leading to improved classification accuracy and more robust detection of early OA indicators.

Enhancing the Discriminative Shape-Texture Convolutional Neural Networks (DST-CNN) model

Exploring Alternative Loss Functions: The proposed discriminative loss function has been effective in enforcing separation between OA and non-OA features. However, there is room for exploring alternative loss functions or regularization techniques that can further enhance the discriminative power of the model. Adaptive or dynamic loss functions that can adaptively adjust the emphasis on different classes or features could be investigated to improve classification performance.

Enhancing Texture Analysis: Texture analysis is crucial for detecting early symptoms and bone changes in OA. We can improve the texture analysis module of the DST-CNN by incorporating more advanced techniques inspired from style transfer, texture synthesis, or texture inpainting. These techniques can help preserve or restore the fine-grained details and textures of the input images. These enhancements could make DST-CNN a more accurate and reliable classifier for OA severity assessment.

Additional potential future directions

Incorporating multi-modal data: In this thesis, the focus has primarily been on using radiographic image data for OA classification. However, OA is a complex condition that can involve various modalities

such as textual reports, patient demographics, or genetic data. Future work can explore the integration of multi-modal data sources to improve the accuracy and robustness of the classification models. This could involve developing fusion techniques or incorporating Transformer attention mechanism to leverage information from different modalities.

Domain adaptation: Applying deep learning models trained on natural image dataset like ImageNet to another medical dataset like OAI and MOST with different characteristics often leads to performance degradation due to domain shift. Investigating domain adaptation methods can help bridge the gap between different datasets and enhance the generalizability of the proposed deep learning models for OA classification. This includes exploring techniques such as unsupervised domain adaptation, domain-specific feature extraction, and data augmentation strategies specific to target domains.

6.3 List of publications

The contributions reported in this thesis was the subject of the following publications :

Journal Articles

- Yassine Nasser, Rachid Jennane, Aladine Chetouani, Eric Lespessailles, Mohammed El Hassouni. "Discriminative Regularized Auto-Encoder for Early Detection of Knee OsteoArthritis : Data from the Osteoarthritis Initiative", IEEE TRANSACTIONS ON MEDICAL IMAGING, vol. 39, no. 9, pp. 2976-2984, Sept. 2020. [Nasser et al., 2020]

- Yassine Nasser, Mohammed El Hassouni, Didier Hans, Rachid Jennane. "A Discriminative Shape-Texture Convolutional Neural Network for Early Diagnosis of Knee Osteoarthritis from X-ray images", Physical and Engineering Sciences in Medicine, 827–837 (2023). [Nasser et al., 2023]

Conference Papers

- Nasser, Yassine, Mohammed El Hassouni, and Rachid Jennane. "Discriminative Deep Neural Network for Predicting Knee OsteoArthritis in Early Stage." International Workshop on PRedictive Intelligence In MEdicine MICCAI. Springer, Cham, 2022. [Nasser et al., 2022]

- Yassine Nasser, Mohammed El Hassouni, Abdelbasset Brahim, Hechmi Toumi, Mohamed Hedi Bedoui, Eric Lespessailles, Rachid Jennane, " Diagnosis of osteoporosis disease from bone X-ray images with Stacked Sparse Autoencoder and SVM classifier", 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP'2017), 2017, pp. 1-5. [Nasser et al., 2017]

Others

- Yassine Nasser, Abdessamad Tafraouti, Mohammed El Hassouni, Hechmi Toumi, Mohamed Hedi Bedoui, Eric Lespessailles, Rachid Jennane, " Diagnostic de l'ostéoporose en utilisant l'auto-encodeur profond avec SVM ", 8ème WORKSHOP AMINA "Applications Médicales de l'Informatique : Nouvelles Approches", le 17, 18 et 19 Novembre 2016 à Monastir, Tunisie. [Nasser et al., 2016]

Bibliography

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

[Andrearczyk and Whelan, 2016] Andrearczyk, V. and Whelan, P. F. (2016). Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69.

[Antony et al., 2017] Antony, J., McGuinness, K., Moran, K., and O'Connor, N. E. (2017). Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International conference on machine learning and data mining in pattern recognition*, pages 376–390. Springer.

[Bijlsma et al., 2011] Bijlsma, J. W., Berenbaum, F., and Lafeber, F. P. (2011). Osteoarthritis: an update with relevance for clinical practice. *The Lancet*, 377(9783):2115–2126.

[Bishop and Nasrabadi, 2006] Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

[Bourlard and Kamp, 1988] Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294.

[Brahim et al., 2019] Brahim, A., Jennane, R., Riad, R., Janvier, T., Khedher, L., Toumi, H., and Lespessailles, E. (2019). A decision support tool for early detection of knee osteoarthritis using x-ray imaging and machine learning: Data from the osteoarthritis initiative. *Computerized Medical Imaging and Graphics*, 73:11–18.

[Cai et al., 2018] Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., and Tong, Y. (2018). Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

[Chen et al., 2019] Chen, P., Gao, L., Shi, X., Allen, K., and Yang, L. (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92.

[Cheng et al., 2018] Cheng, G., Yang, C., Yao, X., Guo, L., and Han, J. (2018). When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE transactions on geoscience and remote sensing*, 56(5):2811–2821.

[Cibrián Uhalte et al., 2017] Cibrián Uhalte, E., Wilkinson, J. M., Southam, L., and Zeggini, E. (2017). Pathways to understanding the genomic aetiology of osteoarthritis. *Human molecular genetics*, 26(R2):R193–R201.

[Cimpoi et al., 2015] Cimpoi, M., Maji, S., and Vedaldi, A. (2015). Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836.

[Cootes et al., 2004] Cootes, T. F., Taylor, C. J., et al. (2004). Statistical models of appearance for computer vision.

[Cross et al., 2014] Cross, M., Smith, E., Hoy, D., Nolte, S., Ackerman, I., Fransen, M., Bridgett, L., Williams, S., Guillemin, F., Hill, C. L., et al. (2014). The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Annals of the rheumatic diseases*, 73(7):1323–1330.

[Culvenor et al., 2015] Culvenor, A. G., Engen, C. N., Øiestad, B. E., Engebretsen, L., and Risberg, M. A. (2015). Defining the presence of radiographic knee osteoarthritis: a comparison between the kellgren and lawrence system and oarsi atlas criteria. *Knee Surgery, Sports Traumatology, Arthroscopy*, 23(12):3532–3539.

[Dai et al., 2017] Dai, X., Yue-Hei Ng, J., and Davis, L. S. (2017). Fason: First and second order information fusion network for texture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7352–7360.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[Durand, 2017] Durand, T. (2017). deep_archi_latex.

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[Fukushima and Miyake, 1982] Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

[Gao et al., 2016] Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326.

[Gatys et al., 2015] Gatys, L., Ecker, A. S., and Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28.

[Goldberger et al., 2004] Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R. (2004). Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

[Goldring, 2009] Goldring, S. R. (2009). Role of bone in osteoarthritis pathogenesis. *Medical Clinics of North America*, 93(1):25–35.

[Haverkamp et al., 2011] Haverkamp, D. J., Schiphof, D., Bierma-Zeinstra, S. M., Weinans, H., and Waarsing, J. H. (2011). Variation in joint shape of osteoarthritic knees. *Arthritis & Rheumatism*, 63(11):3401–3407.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Heidari, 2011] Heidari, B. (2011). Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part i. *Caspian journal of internal medicine*, 2(2):205.

[Hirvasniemi et al., 2014] Hirvasniemi, J., Thevenot, J., Immonen, V., Liikavainio, T., Pulkkinen, P., Jämsä, T., Arokoski, J., and Saarakkala, S. (2014). Quantification of differences in bone texture from plain radiographs in knees with and without osteoarthritis. *Osteoarthritis and cartilage*, 22(10):1724–1731.

[Hladůvka et al., 2017] Hladůvka, J., Phuong, B. T. M., Ljuhar, R., Ljuhar, D., Rodrigues, A. M., Branco, J. C., and Canhão, H. (2017). Femoral rois and entropy for texture-based detection of osteoarthritis from high-resolution knee radiographs. *arXiv preprint arXiv:1703.09296*.

[Hofman et al., 2009] Hofman, A., Breteler, M., van Duijn, C. M., Janssen, H. L., Krestin, G. P., Kuipers, E. J., Stricker, B. H. C., Tiemeier, H., Uitterlinden, A. G., Vingerling, J. R., et al. (2009). The

rotterdam study: 2010 objectives and design update. *European journal of epidemiology*, 24(9):553–572.

[Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[Iorio and Healy, 2003] Iorio, R. and Healy, W. L. (2003). Unicompartmental arthritis of the knee. *JBJS*, 85(7):1351–1364.

[Janvier et al., 2017] Janvier, T., Jennane, R., Toumi, H., and Lespessailles, E. (2017). Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 25(12):2047–2054.

[Kellgren and Lawrence, 1957] Kellgren, J. H. and Lawrence, J. (1957). Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Konwer et al., 2022] Konwer, A., Xu, X., Bae, J., Chen, C., and Prasanna, P. (2022). Temporal context matters: enhancing single image prediction with disease progression representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18824–18835.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[LeCun et al., 2012] LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

[Ledingham et al., 1993] Ledingham, J., Regan, M., Jones, A., and Doherty, M. (1993). Radiographic patterns and associations of osteoarthritis of the knee in patients referred to hospital. *Annals of the rheumatic diseases*, 52(7):520–526.

[Lee et al., 2021] Lee, L. S., Chan, P. K., Fung, W. C., Chan, V. W. K., Yan, C. H., and Chiu, K. Y. (2021). Imaging of knee osteoarthritis: A review of current evidence and clinical guidelines. *Musculoskeletal Care*, 19(3):363–374.

[Lim and Lau, 2011] Lim, K. and Lau, C. S. (2011). Perception is everything: Oa is exciting. *International journal of rheumatic diseases*, 14(2):111–112.

[Lin et al., 2017] Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2017). Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1309–1322.

[Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

[Liu et al., 2019] Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., and Pietikäinen, M. (2019). From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109.

[Lynch et al., 1991a] Lynch, J., Hawkes, D., and Buckland-Wright, J. (1991a). Analysis of texture in macroradiographs of osteoarthritic knees, using the fractal signature. *Physics in Medicine & Biology*, 36(6):709.

[Lynch et al., 1991b] Lynch, J., Hawkes, D., and Buckland-Wright, J. (1991b). A robust and accurate method for calculating the fractal signature of texture in macroradiographs of osteoarthritic knees. *Medical Informatics*, 16(2):241–251.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Messent et al., 2005] Messent, E., Buckland-Wright, J., and Blake, G. (2005). Fractal analysis of trabecular bone in knee osteoarthritis (oa) is a more sensitive marker of disease status than bone mineral density (bmd). *Calcified tissue international*, 76(6):419–425.

[Mika et al., 1999] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee.

[Minciullo and Cootes, 2016] Minciullo, L. and Cootes, T. (2016). Fully automated shape analysis for detection of osteoarthritis from lateral knee radiographs. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 3787–3791. IEEE.

[Minciullo et al., 2017] Minciullo, L., Thomson, J., and Cootes, T. F. (2017). Combination of lateral and pa view radiographs to study development of knee oa and associated pain. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 255–261. SPIE.

[Moody, 1991] Moody, J. (1991). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. *Advances in neural information processing systems*, 4.

[Nasser et al., 2017] Nasser, Y., El Hassouni, M., Brahim, A., Toumi, H., Lespessailles, E., and Jennane, R. (2017). Diagnosis of osteoporosis disease from bone x-ray images with stacked sparse autoencoder and svm classifier. In *2017 international conference on advanced technologies for signal and image processing (ATSIP)*, pages 1–5. IEEE.

[Nasser et al., 2023] Nasser, Y., El Hassouni, M., Hans, D., and Jennane, R. (2023). A discriminative shape-texture convolutional neural network for early diagnosis of knee osteoarthritis from x-ray images. *Physical and Engineering Sciences in Medicine*, pages 1–11.

[Nasser et al., 2022] Nasser, Y., Hassouni, M. E., and Jennane, R. (2022). Discriminative deep neural network for predicting knee osteoarthritis in early stage. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 126–136. Springer.

[Nasser et al., 2020] Nasser, Y., Jennane, R., Chetouani, A., Lespessailles, E., and El Hassouni, M. (2020). Discriminative regularized auto-encoder for early detection of knee osteoarthritis: data from the osteoarthritis initiative. *IEEE transactions on medical imaging*, 39(9):2976–2984.

[Nasser et al., 2016] Nasser, Y., Tafraouti, A., El, M., Toumi, H., Bedoui, M. H., Lespessailles, E., and Jennane, R. (2016). Diagnostic de l'ostéoporose en utilisant l'auto-encodeur profond avec svm.

[Ng et al., 2011] Ng, A. et al. (2011). Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19.

[Nguyen et al., 2020] Nguyen, H. H., Saarakkala, S., Blaschko, M. B., and Tiulpin, A. (2020). Semixup: in-and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs. *IEEE Transactions on Medical Imaging*, 39(12):4346–4356.

[Norman et al., 2019] Norman, B., Pedoia, V., Noworolski, A., Link, T. M., and Majumdar, S. (2019). Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *Journal of digital imaging*, 32(3):471–477.

[Olah et al., 2017] Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11):e7.

[Peterfy et al., 2008] Peterfy, C. G., Schneider, E., and Nevitt, M. (2008). The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage*, 16(12):1433–1441.

[Radin and Rose, 1986] Radin, E. L. and Rose, R. M. (1986). Role of subchondral bone in the initiation and progression of cartilage damage. *Clinical orthopaedics and related research*, (213):34–40.

[Riad et al., 2018] Riad, R., Jennane, R., Brahim, A., Janvier, T., Toumi, H., and Lespessailles, E. (2018). Texture analysis using complex wavelet decomposition for knee osteoarthritis detection: Data from the osteoarthritis initiative. *Computers & Electrical Engineering*, 68:181–191.

[Ruder, 2016] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[Segal et al., 2013] Segal, N. A., Nevitt, M. C., Gross, K. D., Hietpas, J., Glass, N. A., Lewis, C. E., and Torner, J. C. (2013). The multicenter osteoarthritis study: opportunities for rehabilitation research. *PM&R*, 5(8):647–654.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[Shamir et al., 2008] Shamir, L., Ling, S. M., Scott, W. W., Bos, A., Orlov, N., Macura, T. J., Eckley, D. M., Ferrucci, L., and Goldberg, I. G. (2008). Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415.

[Shin et al., 2012] Shin, H.-C., Orton, M. R., Collins, D. J., Doran, S. J., and Leach, M. O. (2012). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1930–1943.

[Shock, 1984] Shock, N. W. (1984). *Normal human aging: The Baltimore longitudinal study of aging*. Number 84. US Department of Health and Human Services, Public Health Service, National ....

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Sohn, 2016] Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

[Springenberg et al., 2014] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Tajbakhsh et al., 2016] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

[Tartaglione et al., 2018] Tartaglione, E., Lepsøy, S., Fiandrotti, A., and Francini, G. (2018). Learning sparse neural networks via sensitivity-driven regularization. *Advances in neural information processing systems*, 31.

[Thomson et al., 2015] Thomson, J., O'Neill, T., Felson, D., and Cootes, T. (2015). Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 127–134. Springer.

[Tiulpin et al., 2018] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., and Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1–10.

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Wen et al., 2016] Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer.

[Woloszynski et al., 2010] Woloszynski, T., Podsiadlo, P., Stachowiak, G., and Kurzynski, M. (2010). A signature dissimilarity measure for trabecular bone texture in knee radiographs. *Medical physics*, 37(5):2030–2042.

[Woloszynski et al., 2012] Woloszynski, T., Podsiadlo, P., Stachowiak, G., Kurzynski, M., Lohmander, L., and Englund, M. (2012). Prediction of progression of radiographic knee osteoarthritis using tibial trabecular bone texture. *Arthritis & Rheumatism*, 64(3):688–695.

[Xu et al., 2015] Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., and Madabhushi, A. (2015). Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

[Zhang et al., 2017] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

A

Appendix

A.1 CNN main building blocks

Convolutional neural networks (CNNs) are a type of deep learning model that have achieved state-of-the-art performance on a wide range of tasks, including image classification, object detection, and language translation. CNNs are particularly well-suited to tasks involving structured data, such as images, because they are able to learn hierarchical representations of the data (see Section 3.3.3). In this appendix, we will describe the main building blocks of CNNs, including convolutional layers, pooling layers, fully connected layers, and activation functions.

Convolutional layer

The core building block of a CNN is the convolutional layer, which consists of a set of learnable filters that are applied to the input data using a sliding window approach. More specificly, convolutional layers apply a convolution operation to the input data, which involves sliding a small window, or kernel, over the input and performing an element-wise multiplication followed by summing up the results. This process is repeated for all locations in the input image, resulting in a feature map that encodes the presence of

important patterns in the input image. The kernel is typically initialized with small random weights and is updated during training through backpropagation.

Given an input volume of size $WxHxD$, a convolutional layer with F filters of size $KxKxD$ will produce an output volume of size $W'xH'xF$, where W' and H' are the width and height of the output volume and F is the number of filters. The output volume is obtained by applying the convolutional filters to the input volume and summing the results. The convolutional operation is defined by the following equation:

$$O(i,j,k) = \sum_{m}\sum_{n}\sum_{d} X(i+m, j+n, d) * W(m,n,d,k) + b(k) \tag{A.1}$$

where O is the output volume, X is the input volume, W is the set of filters, and b is the bias term. The indices i, j, and k index the width, height, and depth dimensions of the output volume, respectively, while the indices m, n, and d index the corresponding dimensions of the filters.

By stacking multiple convolutional layers, a CNN is able to learn increasingly complex patterns in the input image, allowing it to recognize objects and their relationships to one another. The resulting feature maps are then typically fed into one or more pooling layers.

Pooling layer

Another important building block of CNNs is the pooling layer, which is used to down-sample the feature maps produced by the convolutional layers. This helps to reduce the computational complexity of the network and makes it more robust to small translations in the input image. There are several types of pooling, including max pooling, average pooling, and global pooling. Max pooling takes the maximum value within a local window as the output, while average pooling takes the average value.

Given an input volume of size $WxHxD$, a pooling layer with pooling size KxK will produce an output volume of size $W'xH'xD$, where W' and H' are the width and height of the output volume and D is the depth of the input volume. The pooling operation is defined by the following equation for max pooling:

$$O(i,j,k) = max(X(is : is+K-1, js : js+K-1, k)) \tag{A.2}$$

where O is the output volume, X is the input volume, and s is the stride size. The indices i and j index the width and height dimensions of the output volume, respectively, while the index k indexes the depth dimension of the input volume.

The global pooling performs pooling over the entire input and is often used as a way to reduce the number of parameters in the model.

Fully connected layer

Fully connected layers, also known as dense layers, are another key component of CNNs, which are similar to those found in traditional artificial neural networks. These layers take the flattened output of the convolutional or the global pooling layer and performs a linear combination of the inputs. These layers connect every neuron in one layer to every neuron in the next layer via a set of set of weights and biases, allowing the CNN to learn a global representation of the data.

Given an input volume of size $WxHxD$, a fully connected layer with K units will produce an output volume of size $1x1xK$, where K is the number of units in the fully connected layer. The fully connected operation is defined by the following equation:

$$O(k) = \sum_{w,h,d} X(w,h,d) * W(w,h,d,k) + b(k) \tag{A.3}$$

where O is the output volume, X is the input volume, W is the set of weights, and b is the bias term. The index k indexes the units in the output volume, while the indices w, h, and d index the corresponding dimensions of the input volume.

The output of the fully connected layer is a vector of size K, where each element represents the strength of the connection between the input volume and a particular unit in the fully connected layer. These connections (weights and bias) are learned during the training process, allowing the fully connected layer to combine the features learned by the previous layers make the final prediction in a way that is relevant to the task at hand.

Activation function

The activation functions are used in CNNs to introduce non-linearity into the model. Common activation functions include the rectified linear unit (ReLU), sigmoid, and hyperbolic tangent (tanh). Activation functions are typically applied element-wise to the output of a convolutional or fully connected layer.

CNNs also often include additional components such as skip connections and normalization layers. Skip connections allow the model to directly pass the input data or a transformed version of it to the output, allowing the model to better preserve information throughout the network. Normalization layers, such as batch normalization, help to stabilize the training process by normalizing the activations of the layers.

In summary, CNNs are powerful deep learning models that consist of a series of convolutional, pooling, fully connected, activation functions, and optional skip and normalization layers. These building blocks work together to learn and extract important hierarchical representations from raw data, ultimately leading to accurate predictions. In the following sections, we will discuss some of the most popular CNN architectures.

A.1.1 Most popular CNN architectures

In this chapter, we will survey the most popular CNN architectures and delve into their key features, capabilities, and the mathematics behind their modularization. There are numerous CNN architectures that have been developed over the past decade, each with its own strengths and weaknesses. We will focus on some of the most widely-used CNN architectures, including VGGNet, ResNet, Inception, and MobileNet, and discuss their practical applications and performance characteristics.

A.1.1.A LeNet-5

LeNet, developed by Yann LeCun et al. in 1998 [LeCun et al., 1998], is a convolutional neural network (CNN) that was one of the first to be widely used. It is a relatively small network consisting of a few convolutional and fully-connected layers. LeNet was designed to recognize handwritten digits and was used to demonstrate the feasibility of using neural networks for image classification tasks. It consists of two sets of convolutional and pooling layers, followed by fully connected layers. The convolutional layers are responsible for learning local patterns in the input data, while the pooling layers reduce the dimensionality of the data by taking the maximum value within a window. The fully connected layers then process the high-level features extracted by the convolutional and pooling layers to make the final classification. LeNet introduced many key concepts that are still used in modern CNNs, such as the use of convolutional filters to extract local features and the use of pooling to reduce the spatial size of the input and increase the model's ability to generalize. The LeNet-5 architecture is illustrate in Figure A.1:

- The first layer of LeNet, $C1$, is a convolutional layer with 6 filters of size $5 \times 5 \times 1$, where 1 is the depth of the input volume (since the input is a grayscale image). The output of this layer is a volume of size $28 \times 28 \times 6$, where 28×28 corresponding to the size of the input image and 6 is the number of filters. The convolutional operation computed as defined in the Eq. A.1.
- The second layer of LeNet, $S2$, is a pooling layer with a pooling size of 2×2 and a stride of 2. The output of this layer is a volume of size $14 \times 14 \times 6$, where 14×14 is the size of the pooling region

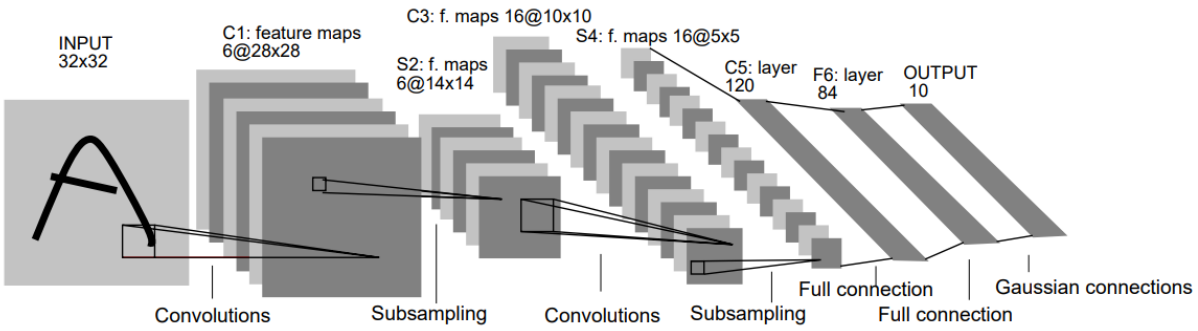


Figure A.1: Architecture of LeNet-5 proposed in [LeCun et al., 1989] for digits recognition.

and 6 is the depth of the input volume. The pooling operation is computed by the equation A.2.

- The third layer of LeNet, $C3$, is another convolutional layer with 16 filters of size $5 \times 5 \times 6$, where 6 is the depth of the input volume. The output of this layer is a volume of size $10 \times 10 \times 16$, where 10×10 is the size of the convolutional region and 16 is the number of filters.
- The fourth layer, $S4$, is another pooling layer with a pooling size of 2×2 and a stride of 2. The output of this layer is a volume of size $5 \times 5 \times 16$, where 5×5 is the size of the pooling region and 16 is the depth of the input volume.
- The fifth layer of LeNet, $C5$, is a convolutional layer with 120 feature maps. Each unit is connected to a 5×5 of all 16 $S4$'s outputs. Thus, the size of $C5$'s feature maps is 1×1.
- The sixth layer of LeNet, $F6$, is a fully connected layer with 84 units. The fully connected operation is can be computed by using the equation A.3.
- The final layer of LeNet, $OUTPUT$, is another fully connected layer with 10 units (one for each digit). Thus, the output of this layer is a vector of size 10, where each element represents the probability of the input image belonging to one of the 10 digit classes.

LeNet is trained using a supervised learning approach, where the network is provided with a labeled dataset of images and their corresponding labels. The network learns to predict the labels of new, unseen images by minimizing a loss function that measures the difference between the predicted labels and the ground truth labels. To optimize the network's weights and biases, an optimization algorithm such as stochastic gradient descent is used to adjust the parameters of the network in a way that reduces the loss. During training, the network is presented with the training set multiple times and the optimization algorithm is used to update the network's weights and biases in an attempt to minimize the loss and improve the network's performance. Once the network is trained, we can use it to make predictions on new, unseen images by feeding the images through the network and selecting the class with the highest predicted probability as the final prediction.

To sum up, LeNet is a pioneering CNN architecture that has had a significant impact on the field of deep learning. Its modular design, consisting of convolutional and pooling layers followed by fully connected layers, has served as the foundation for many subsequent CNN architectures. Its formulation mathematics, including the convolutional and fully connected operations, are essential for understanding how LeNet and other CNNs learn to recognize patterns in image data.

A.1.1.B AlexNet

AlexNet is a convolutional neural network (CNN) architecture developed by Alex Krizhevsky et al. in 2012 [Krizhevsky et al., 2012]. It was the first CNN to achieve a top-5 error rate of less than 15% on

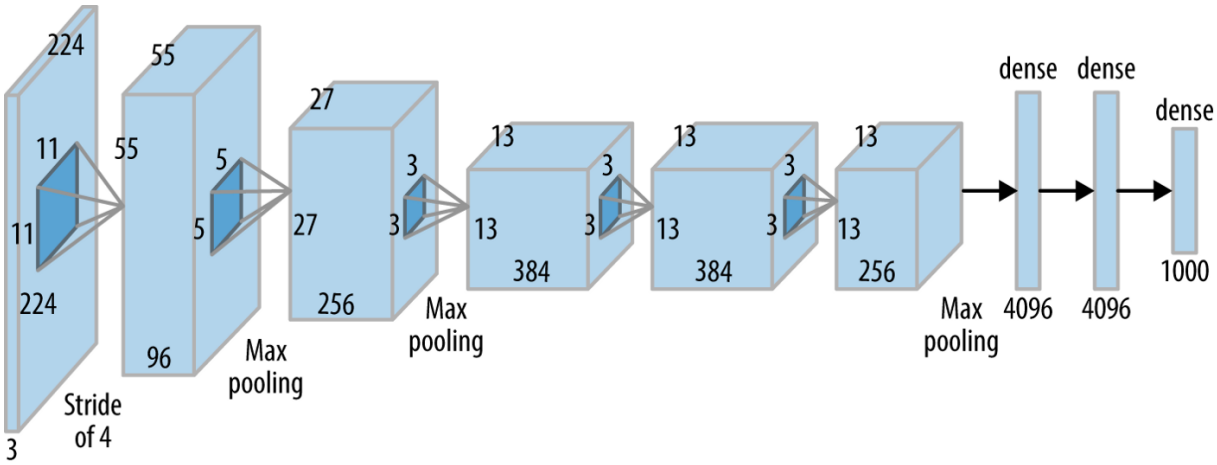


Figure A.2: An illustration of the architecture of AlexNet proposed in [Krizhevsky et al., 2012].

the ImageNet dataset, which contains over 1 million images and 1000 classes. AlexNet consists of eight layers: five convolutional layers, two fully connected layers, and one output layer. An illustration of the architecture of AlexNet is shown in Figure A.2:

- The first layer of AlexNet is a convolutional layer with 96 filters of size $11 \times 11 \times 3$, where 3 is the depth of the input volume (since the input is a color image). The output of this layer is a volume of size $55 \times 55 \times 96$, where 55×55 is the size of the convolutional region and 96 is the number of filters.

- The second layer of AlexNet is a pooling layer with a pooling size of 3×3 and a stride of 2. The output of this layer is a volume of size $27 \times 27 \times 96$, where the size of the pooling region is 3×3 and the stride is 2.

- The remaining layers of AlexNet are similar to the first two layers, consisting of additional convolutional and pooling layers followed by fully connected layers. The final layer is a 1000-way softmax layer, which produces a probability distribution over the 1000 possible classes. The fully connected operation is defined in the same way as before.

AlexNet introduced the use of rectified linear unit (ReLU) as an activation function, which has since become a standard in CNNs due to its ability to improve the training speed, and Dropout as a regularization technique to reduce overfitting. AlexNet was also the first CNN to demonstrate that deep learning could significantly outperform traditional approaches to image classification. It won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, significantly surpassing the performance of the previous state-of-the-art methods.

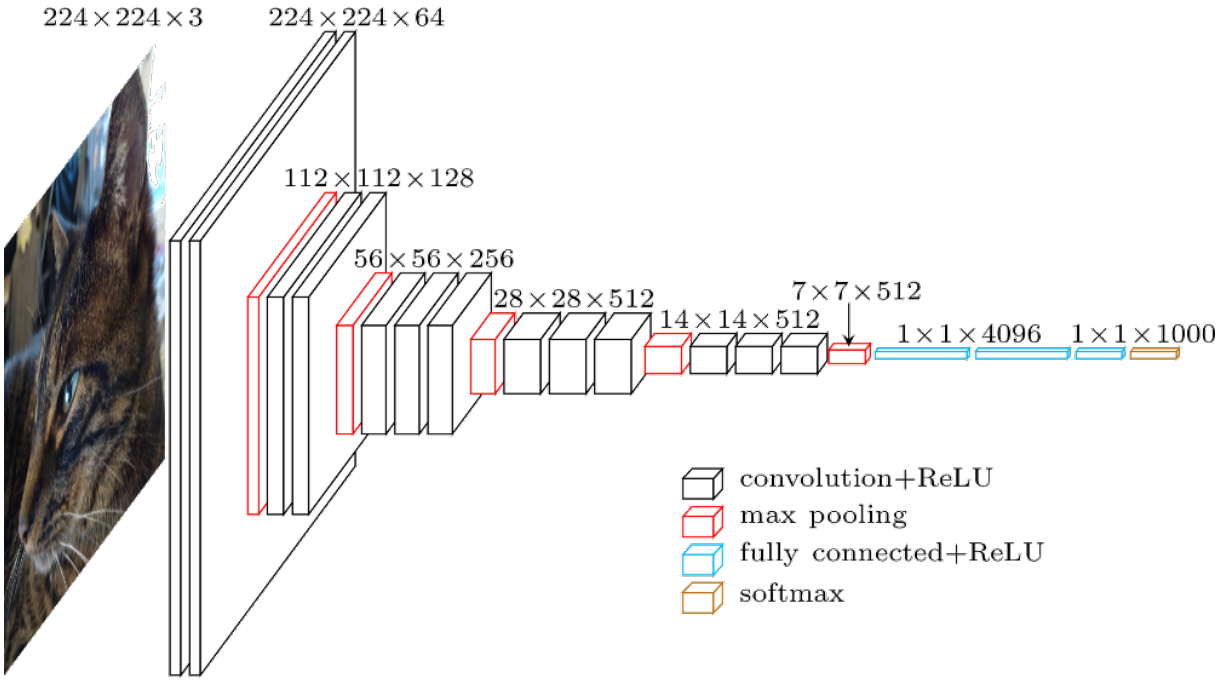


Figure A.3: Illustration of an architecture of a typical VGG16. Figure by [Durand, 2017]

.

A.1.1.C VGGNet

VGGNet is a CNN architecture developed by Karen Simonyan and Andrew Zisserman in 2014 [Simonyan and Zisserman, 2014]. It is known for its simplicity and high performance on a wide range of image classification tasks. VGGNet consists of a stack of convolutional layers with small filters, followed by a number of fully connected layers. One key feature of VGGNet is its use of a consistent number of convolutional layers (16 or 19) across all of its architectures, which allows for easy comparison of different models. A typical architecture of VGGNet is illustrated in Figure A.3.

The convolutional layers in VGGNet use a small filter size of 3×3, which allows the network to learn a large number of filters and therefore capture more detailed features from the input data. After each convolutional layer, VGGNet includes a max-pooling layer with a pooling size of 2×2 and a stride of 2. The output of the pooling layer is a volume of size $W/2 \times H/2 \times D$, where W, H, and D are the width, height, and depth of the input volume, respectively.

The fully connected layers in VGGNet process the high-level features extracted by the convolutional and pooling layers and produce the final output of the network. The fully connected operation is defined in the same way as in traditional neural networks.

VGGNet has achieved state-of-the-art results on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and other benchmarks. It has also been used in a variety of applications, including object detection and segmentation. The small convolutional filters allow VGGNet to learn more detailed

features from the input data, and the deep structure of the network allows it to capture complex patterns in the data.

A.1.1.D Inception

The Inception network, also known as GoogLeNet, is a deep convolutional neural network architecture that was developed by Google in 2014 for the ImageNet Large Scale Visual Recognition Challenge [Szegedy et al., 2015]. It is called the Inception network because it is inspired by the idea of "inception" or the beginning of something, in this case the beginnings of a deep learning model for image recognition.

One key feature of the Inception network is the use of multiple parallel convolutional layers with different kernel sizes in each module, known as "inception modules". This allows the network to learn a variety of different feature maps at different scales, which helps to improve the accuracy of the model. This is achieved by using a combination of $1x1$, $3x3$, and $5x5$ convolutional filters, and a max pooling layers to reduce the spatial resolution of the feature maps, as well as using a concatenation layer that combines the output feature maps of the convolutional layers (see Fidure A.4).

The 1×1 convolutional filters are used to reduce the number of input channels, which helps to reduce the computational complexity of the network. The 3×3 and 5×5 filters are used to capture larger spatial features, while the max pooling layers are used to reduce the spatial resolution of the feature maps and to introduce some degree of translation invariance. As shown in Figure A.4, the module consists of a series of parallel convolutional layers with different kernel sizes and a concatenation layer that combines the output feature maps of the convolutional layers. The Inception module can be formulated as follows:

$$\text{output} = \text{concatenate}(\text{conv1x1, conv3x3, conv5x5, maxpool3x3})$$

where "conv1x1," "conv3x3," "conv5x5," and "maxpool3x3" are the output feature maps of the 1×1 convolution, 3×3 convolution, 5×5 convolution, and 3×3 max pooling layers, respectively, and

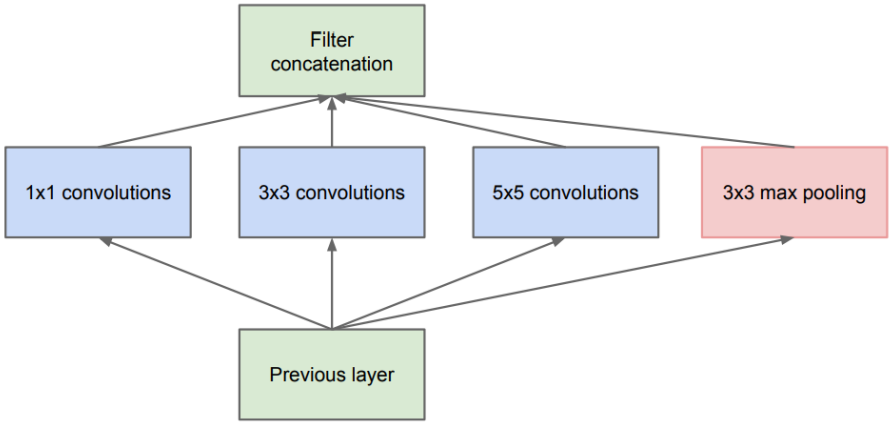


Figure A.4: Inception module presented in [Szegedy et al., 2015]

"concatenate" is a function that combines the feature maps along the channel axis.

Inception network consists of a series of Inception modules stacked together, followed by a global average pooling layer and a fully connected layer. The final output of the network is obtained using a softmax classifier. The network is trained using a variant of stochastic gradient descent called Adam [Kingma and Ba, 2014], which is a first-order gradient-based optimization algorithm that adapts the learning rate for each parameter. The optimization objective is to minimize the loss function $\mathcal{L}$, which measures the difference between the predicted output Y and the true output Y_true:

$$\mathcal{L} = \mathcal{L}(Y, Y_true) \tag{A.4}$$

In addition to the Inception module, the Inception network also utilizes other techniques such as batch normalization and skip connections to improve performance. Batch normalization helps to reduce the internal covariate shift, which is the change in the distribution of network activations due to the change in network parameters during training. Skip connections, also known as residual connections, allow the network to learn identity functions, which can improve the network's ability to learn.

Furthermore, the Inception network also uses global average pooling, which replaces the fully connected layers at the end of the network with a global average pooling layer. This helps to reduce the number of parameters in the network and prevent overfitting.

In summary, the Inception network is a powerful CNN architecture that has achieved strong performance on a wide range of image classification tasks. Its modular design, using inception modules to learn a variety of features at different scales, and its efficient use of parameters make it a popular choice for many image classification applications.

A.1.1.E ResNet

Residual networks (ResNets) are an other type of CNN architecture, that was developed by Microsoft Research in 2015 [He et al., 2016]. They have achieved state-of-the-art performance on many image classification benchmarks. The main advantage of ResNets is that they make it possible to train very deep CNNs without suffering from the vanishing gradient problem.

The vanishing gradient problem is a phenomenon that occurs in deep neural networks, where the gradients of the weights with respect to the loss function become very small as the weights are backpropagated through the network. This makes it difficult for the network to learn and update its weights, as the gradients are too small to make a significant impact.

ResNets address this problem by introducing the concept of residual learning, which allows the network to learn a residual function with respect to the layer input rather than the ground truth. This is achieved through the use of residual connections, which are shortcut connections that skip one or more layers and allow the network to bypass the layers that are causing the vanishing gradient problem.

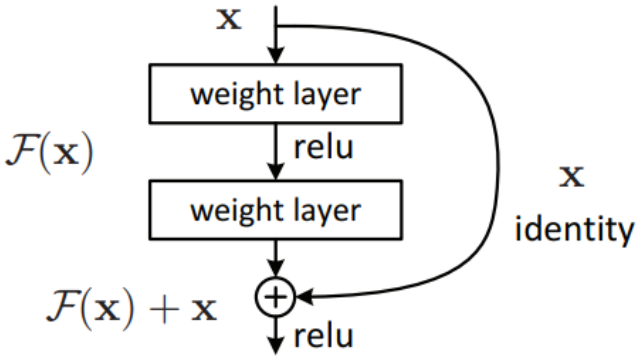


Figure A.5: Residual learning: a building block of the ResNet [Szegedy et al., 2015]

In a traditional CNN, the output of a layer is computed using the following equation:

$$\text{output} = \text{f(input, weights)}$$

where "f" is a non-linear function, such as a rectified linear unit (ReLU), and "weights" are the parameters of the layer. In a ResNet, the output is computed using the follow:

$$output = input + f(input, weights)$$

Mathematically, the residual function can be represented as:

$$F(x) = H(x) - x \tag{A.5}$$

where $F(x)$ is the residual function, $H(x)$ is the output of the layer, and x is the input to the layer (see Figure A.5).

The output of the layer is then computed as:

$$H(x) = F(x) + x \tag{A.6}$$

This means that the output of the layer is the sum of the residual function and the input, rather than the output of the layer itself.

The ResNet architecture consists of a series of convolutional layers, interleaved with residual blocks, followed by a final output layer (Figure A.6). The convolutional layers are used to extract features from the input data, and the residual blocks are used to allow the network to learn more complex features by passing the input data through multiple layers and adding the output of each layer together. The final output layer is used to make predictions based on the learned features.

Overall, the use of residual connections allows the ResNet to learn much deeper architectures without

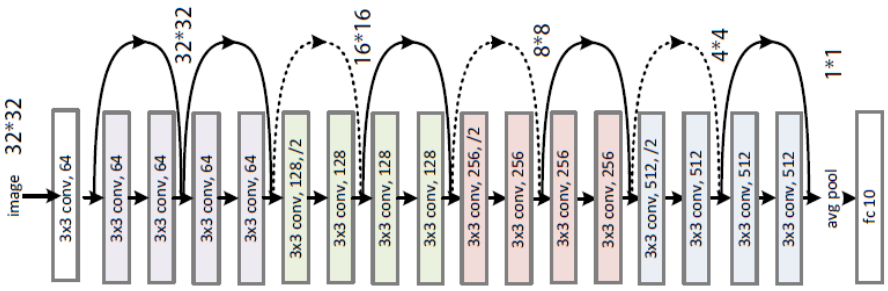


Figure A.6: Illustration of ResNet-18 a typical architecture of ResNet [Szegedy et al., 2015]

suffering from the vanishing gradient problem. ResNets have been very successful and have set new benchmarks in image classification and other tasks.

A.1.1.F DenseNet

DenseNet [Huang et al., 2017] is one of the most popular CNN architectures, it is a densely connected convolutional network pre-trained on ImageNet [Deng et al., 2009].

Let x_l be the output of the l^{th} layer. In traditional CNNs, x_l is computed by applying a nonlinear transformation H_l to previous layer's output x_{l-1}

$$x_l = H_l(x_{l-1}) \tag{A.7}$$

During consecutive convolutions, activation functions, and pooling operations, the network obtains robust semantic features in the top layers. However, fine image details related to texture tend to disappear in the top layers of the network.

Inspired by the main idea of the ResNet learning model [He et al., 2016], which introduces a residual block that sums the identity mapping of the input to the output of a layer, and to improve the information flow between layers, DenseNet proposes a direct connection from any layer to all subsequent layers. Consequently, the l^{th} layer receives the feature maps from all preceding layers as inputs. Thus, it is possible to define the output of the l^{th} layer as:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{A.8}$$

where [...] represents the concatenation operation, $H_l(.)$ is a composite function of the following consecutive operations: Batch Normalization (BN), Rectified Linear Units (ReLU), and a 3×3 Convolution (Conv). We denote such composite function as one layer.

For example, Densenet-121 used in our experiments consists of four dense blocks, each of which has 6, 12, 24, and 16 layers. To reduce the number of feature maps, DenseNet introduces a transition-down block between each two contiguous dense blocks. A transition down layer consists of a batch of normalization

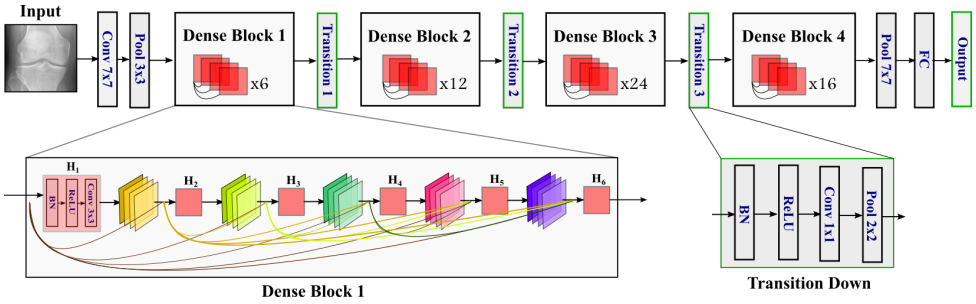


Figure A.7: Architecture of the DenseNet-121 learning model introduced in [He et al., 2016].

followed by a ReLU function, and a 1×1 convolutional layer followed by a 2×2 max pooling. Fig.A.7 provides a schematic overview of the architecture of DenseNet and the composition of each block.

A.1.2 MobileNet

MobileNet was developed by Andrew G. Howard et al. in 2017 [Howard et al., 2017]. It is designed to be used on mobile and embedded devices with limited computational resources, such as smartphones and smartwatches. MobileNet is known for its efficiency and ability to run on low-power devices.

The key innovation of MobileNet is the use of depthwise separable convolutions, which are a type of convolution operation that decomposes the standard convolution operation into two simpler operations: a depthwise convolution and a pointwise convolution. Depthwise convolutions operate on each input channel separately, whereas pointwise convolutions combine the input channels using a 1×1 convolution.

As we saw earlier in Section A.1, the standard convolution operation can be formulated as follows:

$$O[i,j,k] = \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} X[i,j,s] * W[s,t,k] \tag{A.9}$$

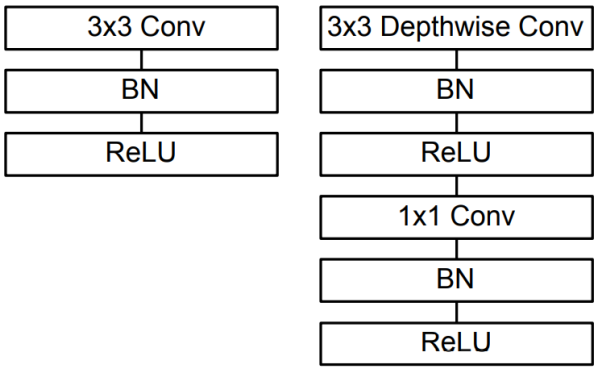


Figure A.8: A comparaison between the standard convolutional layer with batchnorm and ReLU (Left), and the Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU (Right).

where X and W are the input feature map and the convolutional kernel, respectively, and S and T are the spatial dimensions of the kernel. The output feature map, O, is computed by applying the convolutional kernel at every position in the input feature map and summing the results.

The depthwise convolution operation can be formulated as:

$$O[i, j, k] = \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} X[i, j, k] * W[s, t, k] \tag{A.10}$$

The depthwise convolution operation is similar to the standard convolution operation, but it operates on each input channel separately using a separate set of weights. This means that the number of parameters in the depthwise convolution is much smaller than in the standard convolution.

The pointwise convolution operation can be formulated as:

$$O[i, j, k] = \sum_{s=0}^{C-1} X[i, j, s] * W[0, 0, s] \tag{A.11}$$

where C is the number of input channels. The pointwise convolution combines the input channels using a 1×1 convolution with a single set of weights.

The depthwise separable convolution operation is then given by:

$$\text{output[i, j, k]} = \text{depthwise[i, j, k]} + \text{pointwise[i, j, k]}$$

where "depthwise" and "pointwise" are the output feature maps of the depthwise (Eq. A.10) and pointwise (Eq. A.11) convolutions, respectively. Figure A.8, highlight the differences between standard convolution and the depthwise separable convolution.

The use of depthwise separable convolutions significantly reduces the number of parameters and computational requirements of the network, making it more efficient and suitable for use on mobile devices. The architecture of MobileNet consists of several depthwise separable convolutional layers, followed by an average pooling layer and a fully connected layer.

In summary, MobileNet is a CNN architecture developed for use on mobile devices. It is characterized by the use of depthwise separable convolutions, which decompose the standard convolution operation into two simpler operations: a depthwise convolution and a pointwise convolution. This significantly reduces the number of parameters and computational requirements of the network, making it more efficient and suitable for use on low-power devices.

A.1.3 Summary

In this appendix, we introduced convolutional neural networks (CNNs) as a type of neural network architecture that is effective for image classification tasks. We discussed the main components of CNNs,

including convolutional layers, pooling layers, fully connected layers, and activation layers. We also reviewed several popular CNN architectures, such as LeNet-5, AlexNet, VGG, Inception, ResNet, DenseNet, and MobileNet. These architectures have their own unique characteristics and have achieved successful results in various tasks and applications. Overall, CNNs have demonstrated strong performance in image and video recognition tasks and have greatly impacted the field of computer vision.

Yassine NASSER

RÉSEAUX DE NEURONES PROFONDS AMÉLIORÉS POUR LE DIAGNOSTIC PRECOCE DE LA GONARTHROSE

Résumé :
La gonarthrose est une maladie dégénérative du genou qui peut entraîner une douleur et une perte de mobilité. La thèse en question a pour objectif de développer des modèles de d'apprentissage profond pour détecter précocement la gonarthrose à partir d'images radiographiques. Pour cela, une nouvelle architecture, appelée DRAE, basée sur les auto-encodeurs a été introduite. Le but de ce modèle est de séparer les images de genoux sains et arthrosiques en minimisant la distance entre images de même classe (intra-classes) et en maximisant la distance entre images de classes différentes (inter-classes). Ensuite, cette régularisation discriminante a été intégrée à un réseau de neurones convolutionnel (CNN) pour améliorer la détection précoce de la gonarthrose. Ce modèle, appelé DCNN, a été proposé pour analyser à la fois la texture et la forme de l'image. Enfin, le modèle final, appelé DST-CNN, a été proposé pour améliorer l'analyse de la texture et s'adapter aux tâches de classification multi-classes. Le modèle DST-CNN a montré une meilleure performance de classification et bien équilibrée que les modèles de l'état de l'art existants.

Mots clés : Apprentissage profond, Apprentissage des caractéristiques, Auto-Encodeur, réseau de neurones convolutionnel, Régularisation discriminante, Arthrose du genou, Radiographie simple, Rayon X.

ENHANCED DEEP NEURAL NETWORKS FOR EARLY DIAGNOSIS OF KNEE OSTEOARTHRITIS

Abstract :

Knee osteoarthritis is a common cause of physical disability that causes pain and reduced mobility. This thesis aimed to develop deep-learning models to detect knee osteoarthritis at an early stage using radiographic images. A new autoencoder-based architecture called the Discriminative Regularized Auto-Encoder (DRAE) was introduced to achieve this. The DRAE was designed to distinguish between healthy and arthritic knee images by minimizing the distance between images of the same class (intraclass) and maximizing the distance between images of different classes (interclass). This discriminative regularization technique, used in DRAE, was then incorporated into a Convolutional Neural Network (CNN). The resulting model, called Discriminative Convolutional Neural Network (DCNN), was proposed to analyze both the texture and shape of the images, therefore improving the early detection of knee osteoarthritis. Finally, a Discriminative Shape-Texture Convolutional Neural Network (DST-CNN) model was proposed to enhance texture analysis and adapt to the multi-class classification tasks. The DST-CNN model showed better and well-balanced classification performance than current state-of-the-art models.

Keywords : Deep learning, Feature representation learning, Auto- Encoder, CNNs, Discriminative regularization, knee osteoarthritis, plain radiography, X-ray.

Institut Denis Poisson (UMR CNRS 7013),Collegium Sciences et Techniques, Bâtiment de mathématiques, Rue de Chartres B.P. 6759, 45067 Orléans cedex 2 FRANCE