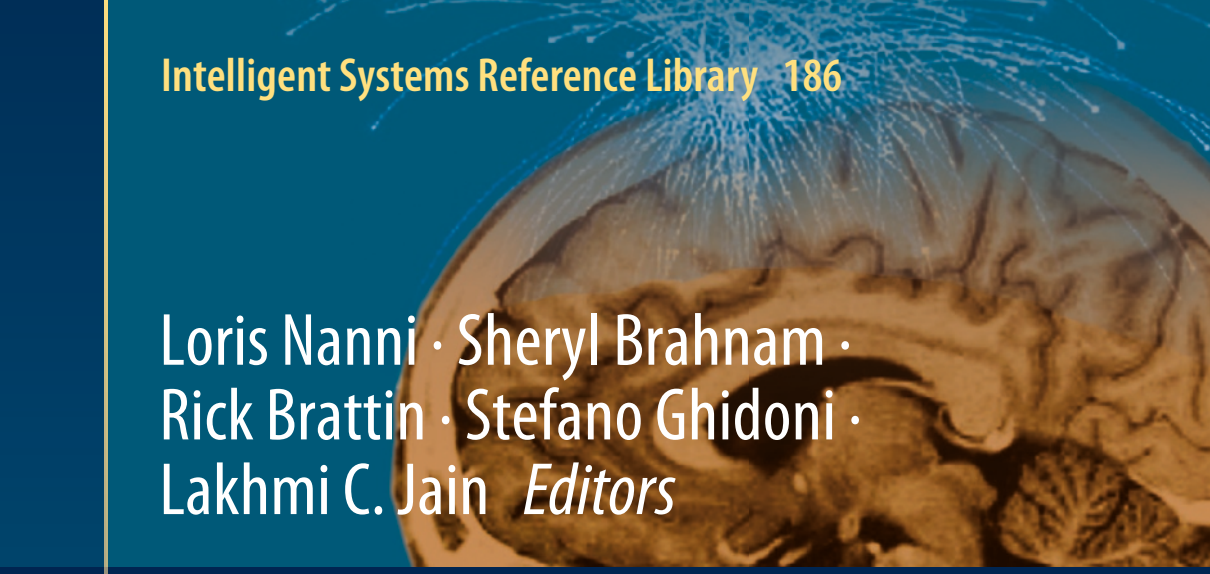
Intelligent Systems Reference Library 186



Loris Nanni · Sheryl Brahnam ·
Rick Brattin · Stefano Ghidoni ·
Lakhmi C. Jain *Editors*

Deep Learners and Deep Learner Descriptors for Medical Applications

Springer

Intelligent Systems Reference Library

Volume 186

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

** Indexing: The books of this series are submitted to ISI Web of Science, SCOPUS, DBLP and Springerlink.

More information about this series at http://www.springer.com/series/8578

Loris Nanni · Sheryl Brahnam ·
Rick Brattin · Stefano Ghidoni ·
Lakhmi C. Jain
Editors

Deep Learners and Deep Learner Descriptors for Medical Applications

Springer

Editors
Loris Nanni
Department of Information Engineering
University of Padova
Padova, Italy

Rick Brattin
Department of Information Technology
and Cybersecurity
Missouri State University
Springfield, MO, USA

Lakhmi C. Jain
University of Technology
Sydney, Australia

Liverpool Hope University
Liverpool, UK

KES International
Shoreham-by-Sea, UK

Sheryl Brahnam
Computer Information Systems
Missouri State University
Springfield, MO, USA

Stefano Ghidoni
Department of Information Engineering
Intelligent Autonomous Systems Laboratory
Padova, Padova, Italy

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-030-42748-1 ISBN 978-3-030-42750-4 (eBook)
https://doi.org/10.1007/978-3-030-42750-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book introduces the reader to current trends using deep learners and deep learner descriptors for medical applications. This volume provides the latest review of the literature and illustrates across a variety of medical image and sound applications the five major ways deep learners can be exploited: (1) by training a deep learner from scratch using data preprocessing, selection, and augmentation techniques to solve imbalances or insufficiencies in the medical data; (2) by utilizing transfer learning from a pre-trained deep learner as a complementary feature extractor for other simpler classifiers, such as the Support Vector Machine (in this method, the resulting learned features are often combined with advanced handcrafted texture features, such as Local Binary Patterns); (3) by fine-tuning one or more pre-trained deep learners on some other unrelated dataset so that it discriminates a novel medical dataset; (4) by fusing different deep learner architectures; and (5) by combining the above methods in a variety of ways to generate more elaborate ensembles. This book will be of value to anyone already engineering deep learners for medical applications, as well as to those interested in learning more about current techniques in this exciting field. Some chapters provide source code that can be used to investigate topics further or to kickstart new projects.

Padova, Italy — Loris Nanni
Springfield, USA — Sheryl Brahnam
Springfield, USA — Rick Brattin
Padova, Italy — Stefano Ghidoni
Sydney, Australia — Lakhmi C. Jain

Contents

Chapter 1
An Introduction to Deep Learners and Deep Learner Descriptors for Medical Applications

Loris Nanni, Sheryl Brahnam, Rick Brattin, Stefano Ghidoni and Lakhmi C. Jain

Abstract This chapter provides an introduction to deep learners and deep learner descriptors for medical applications. A basic outline of the deep learning (DL) process and five methods for exploiting DL with the Convolutional Neural Network (CNN), is presented, as well as a summary of the chapters in this book.

1.1 Introduction

Deep Learning (DL) is one of the best-performing approaches in Artificial Intelligence (AI), a field that was revolutionized when it was first proposed. The main feature of deep learning is its layered structure, with the different layers forming a hierarchy of processing stages ranging from low-level (layers closer to the input), where more generalizable information or descriptors are discovered, to high-level analysis (layers further away from the input), where information flowing through the network represents high-level concepts. What this means is that every layer adds a certain level of abstraction to the overall representation.

Considering medical image interpretation, DL analyzes data through several stages: at a lower level, small image patches are considered, leading to features like edges and texture. Such low-level descriptors are then combined to build more

L. Nanni (✉) · S. Ghidoni
Department of Information Engineering, University of Padua, Via Gradenigo 6, 35131 Padova, Italy
e-mail: loris.nanni@unipd.it

S. Brahnam · R. Brattin
Department of Information Technology and Cybersecurity, Glass Hall, Missouri State University, 901 S. National, Springfield, MO 65804, USA

L. C. Jain
University of Technology, Sydney, Australia
e-mail: jainlakhmi@gmail.com; jainlc2002@yahoo.co.uk

Liverpool Hope University, Liverpool, UK

KES International, Shoreham-by-Sea, UK

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_1

complex representations. At the next level, features like larger image patches and contours are considered. Moving toward the upper levels, closer to the outputs, the elements processed by the network increase in complexity and are extracted from larger areas of the image and from larger sets of the input data.

Recent literature presents a wide variety of articles related to the applications of DL, especially as it pertains to Convolutional Neural Networks (CNN), one of the most powerful deep learners for vision tasks. Broadly, there are five methods for using CNN: (1) training a CNN from scratch using data preprocessing, augmentations, and selection to solve any imbalances or insufficiencies in the data; (2) transfer learning from a pretrained CNN as a complementary feature extractor, where the learned features (sometimes combined with advanced handcrafted image features, such as Local Binary Patterns (LBP) and its variants) are trained on other classifiers, such as the Support Vector Machine (SVM); (3) fine-tuning one or more pretrained CNNs on a novel dataset; (4) fusing different CNN architectures; and (5) combining many of the above methods to generate more elaborate ensembles. All five techniques are exploited in medical applications, texture analysis, and biomedical image and sound processing.

The objective of this book is to bring together key researchers working with DL, as described above, on different medical applications. The majority of chapters in this book focus on CNN and medical images; however, chapters on sound as how DL relates to medical sound applications are also included.

This book is divided into four parts. The first illustrates the use of deep features, namely features that are automatically learned by deep networks. The second focuses on data augmentation, which is often crucial when only small datasets are available for training, not uncommon with many medical imaging datasets due to the expense of collecting such images. The third part presents several applications, demonstrating how wide the adoption of DL-based systems is in the medical field. Finally, the fourth part presents a chapter that combines work in DL with ethical considerations. Ethics is an aspect that is often neglected in papers related to engineering, computer science, and technology, but we consider this to be a crucial aspect of DL, so we have decided to include this chapter to provide a complete picture.

1.2 Outline of This Book

The chapters in this book are organized into four parts:

Part 1: Deep Features and their Fusion
Part 2: Augmentation
Part 3: Medical Applications and Reviews
Part 4: Ethical Considerations.

Part 1, Deep Features and their Fusion, contains four chapters, three of which combine deep features extracted from CNN layers with handcrafted LBP descriptors and their variants. These three chapters focus on the assessment of radiographic

knee osteoarthritis severity, cervical cell classification, and a wide spectrum of other bioimage tasks, including classifying tissue subregions into epithelium, stroma, lymphocytes, and necrosis. The fourth chapter in Part 1 is devoted to presenting methods for extracting deep audio representations for audio-based medical applications. Not all the chapters in part four are focused exclusively on DL features. Chapter 2, for example, also demonstrates the effectiveness of CNN to detect regions of interest (ROI).

Part 2, Augmentation, contains one chapter that reviews data augmentation as it relates specifically to sparse medical data.

Part 3, Medical Applications and Reviews, is made up of three chapters. The first two chapters provide: (1) a systematic review of CNN in detecting gastrointestinal and liver cancers, and (2) a survey of CNNs in the field of intra-operative optical image analysis. The third chapter offers a unique method for protein classification using an ensemble of CNNs trained on 2D multiview images of 3D protein structures generated from a 3D molecular graphics program.

Part 4, Ethical Considerations, contains a chapter that not only reviews CNN architectures, including one developed by the authors, and how these CNNs advance applications in medicine but also examines the ethical implications of producing machines that influence human decision making.

Below we provide a synopsis of each of the nine chapters of this book.

Part 1: Deep Features and Their Fusion

In Chap. 2, Joseph Antony, Kevin McGuinness, Kieran Moran, Noel E. O'Connor discuss and illustrate some of the advantages of using deep features learned from CNN versus handcrafted features by applying CNN to the problem of diagnosing knee osteoarthritis severity from X-ray images using Kellgren and Lawrence (KL) grades. This problem requires that knee joint regions first be detected, which the authors accomplish using a fully CNN (FCN). Once detected, these regions are then labeled and trained from scratch on a CNN to predict the KL grades 0 through 4. In other words, the authors present a fully automatic knee Osteoarthritis diagnostic system using deep learners. Included in this chapter is a description of FCN and how it can be used to detect regions of interest (ROI). The superiority of FCN is compared with template matching and SVM methods to detect knee joints. CNN classification of knee OA severity is compared with SVMs trained on several sets of handcrafted features: LBP and variants, Histogram of oriented gradients (HOG), etc. Features extracted from pre-trained CNNs and the fine-tuning of pre-trained CNNs are also examined and compared with the CNN trained from scratch. Results clearly show that learned features are more capable of handling fine-grained differences between categories than are handcrafted features.

In Chap. 3, Jakob N. Kather, Raquel Bello-Cerezo, Francesco Di Maria, Gabi W. van Pelt, Wilma E. Mesker, Niels Halama, and Francesco Bianconi present an assessment of the effectiveness of deep image features extracted from eight pre-trained CNNs compared with twelve variants of LBP for classifying tissue subregions into epithelium, stroma, lymphocytes, and necrosis. Feature sets are trained on non-parametric nearest-neighbour (1-NN) rule with the cityblock distance measure and

SVM with a radial-basis kernel function. Results of the analysis show that both types of features can be effective for classifying tissue subregions, but features extracted from CNNs exhibit a notable degree of superiority.

In Chap. 4, Loris Nanni, Stefano Ghidoni, Sheryl Brahnam, Shaoxiong Liu, and Ling Zhang present an ensemble of heterogeneous descriptors for bioimage classification. Although the primary focus of this chapter is on cervical cell classification, the ensemble empirically developed in this work is tested on several other bioimage problems with excellent results. This chapter presents methods for building ensembles of CNNs by leveraging the classification power of fine-tuned pretrained CNNs. Several different CNN architectures are assessed by running experiments using different learning rates, batch sizes, and topologies. Features extracted from CNN layers are also trained and combined using a set of SVMs. Finally, handcrafted features, mostly based on variants of LBPs, are extracted from bioimages and are trained separately on SVMs. Results of all methods are combined. This simple ensemble produces a high performing, competitive system, one that outperforms the single best CNN trained specifically on the tested datasets. The authors provide access to all MATLAB source code so that researchers can replicate results.

In Chap. 5, Shahin Amiriparian, Maximilian Schmitt, Sandra Ottl, Maurice Gerczuk, and Björn Schuller introduce a transfer learning approach for extracting audio representation, called DEEP SPECTRUM, which uses CNNs pre-trained on ImageNet for extracting features from audio spectrograms. To solve the data scarcity problem common with medial datasets, the authors propose a deep convolutional generative adversarial network (DCGAN) for unsupervised learning of robust representations from the spectral features along with a recurrent sequence-to-sequence autoencoder for learning fixed-length representations of variable-length audio sequences. The proposed system is evaluated on datasets representing abnormal heart sound classification, snore sound classification, and bipolar disorder recognition. Included in this chapter is a short review of CNN, Recurrent Neural Networks (RNN), autoencoders, and generative adversarial networks (GANs).

Part 2: Augmentation

In Chap. 6, Behnaz Abdollahi, Naofumi Tomita, and Saeed Hassanpour review some of the most beneficial data augmentation methods for image applications, specifically for medical image applications. Some augmentation methods covered include image cropping, flipping, affine transformations, color perturbation, and DL augmentation using synthesized data produced by GANs. This chapter provides many practical guidelines for medical image detection and segmentation, illustrated with some medical applications.

Part 3: Medical Applications and Reviews

In Chap. 7, Samy A. Azer provides an assessment of CNN accuracy in detecting gastrointestinal and liver cancers. Reviewed in this chapter are twenty-two papers that cover esophagus, stomach, pancreas, liver and biliary system and colon cancers as well as some precancerous conditions, liver cirrhosis, and colonic polyps. The chapter also suggests future areas of research and notes both the strengths and limitations of

CNN interpretation for gastrointestinal and liver cancer images. Particularly stressed is the need of control studies for evaluating CNNs compared with expert assessments.

In Chap. 8, Sara Moccia, Luca Romeo, Lucia Migliorelli, Emanuele Frontoni, and Primo Zingaretti provide a survey of CNNs in the field of intra-operative optical image analysis, which builds a patient-specific model that can be used to define a surgical plan that can be updated in the operating room. After an overview of CNN, the survey analyzes about fifty papers published in the last five years that are divided into approaches for optical image segmentation and classification. Included in this chapter is a section that lists and evaluates some publicly available datasets and the metrics used to evaluate algorithm performance.

In Chap. 9, Loris Nanni, Federica Pasquali, Sheryl Brahnam, Alessandra Lumini, and Apostolos Axenopoulos present an ensemble of CNNs that are trained on 2D multiview images of 3D protein structures that were generated from the 3D molecular graphics program Jmol. The images are used to fine-tune two pretrained CNNs, AlexNet and GoogleNet, both of which were trained on ImageNet. The ensemble is then tested on two datasets demonstrating the usefulness of this unique approach to 3D protein classification as compared with the state-of-the-art. The authors provide access to all MATLAB source code so that researchers can replicate results.

Part 4: Ethical Considerations

In Chap. 10, O. Lucia Quintero Montoya and Juan Guillermo Paniagua provide a refreshing critical review of AI as it assumes more of the creative and learning capacities of human beings, examine the ethical implications of producing machines that influence human decision making, review CNN architectures and how they advance applications in medicine, and then finally present their own CNN architecture called Convolutional Laguerre Gauss Network. Of interest in this work is how it interweaves ethical concerns with advancements in the state-of-the-art, stressing thereby the necessity of researchers in the area of AI to consider the ethical implications of their work.

Part I
Deep Features and Their Fusion

Chapter 2
Feature Learning to Automatically Assess Radiographic Knee Osteoarthritis Severity

Joseph Antony, Kevin McGuinness, Kieran Moran and Noel E. O'Connor

Abstract Feature learning refers to techniques that learn to transform raw data input into an effective representation for further higher-level processing in many computer vision tasks. This chapter presents the investigations and the results of feature learning using convolutional neural networks to automatically assess knee osteoarthritis (OA) severity and the associated clinical and diagnostic features of knee OA from radiographs (X-ray images). Also, this chapter demonstrates that feature learning in a supervised manner is more effective than using conventional handcrafted features for automatic detection of knee joints and fine-grained knee OA image classification. In the general machine learning approach to automatically assess knee OA severity, the first step is to localize the region of interest that is to detect and extract the knee joint regions from the radiographs, and the next step is to classify the localized knee joints based on a radiographic classification scheme such as Kellgren and Lawrence grades. First, the existing approaches for detecting (or localizing) the knee joint regions based on handcrafted features are reviewed and outlined in this chapter. Next, three new approaches are introduced: (1) to automatically detect the knee joint region using a fully convolutional network, (2) to automatically assess the radiographic knee OA using CNNs trained from scratch for classification and regression of knee joint images to predict KL grades in ordinal and continuous scales, and (3) to quantify the knee OA severity optimizing a weighted ratio of two loss functions: categorical cross entropy and mean-squared error using multi-objective convolutional learning. The results from these methods show progressive improvement in the overall quantification of the knee OA severity. Two public datasets: the OAI and the MOST are used to evaluate the approaches with promising results that outperform existing approaches. In summary, this work primarily contributes to the field of automated methods for localization (automatic detection) and quantification (image classification) of radiographic knee OA.

J. Antony (✉) · K. McGuinness · K. Moran · N. E. O'Connor
Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
e-mail: joseph.antony@insight-centre.org

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_2

Keywords Feature learning · Handcrafted features · Convolutional neural networks · Kellgren and Lawrence grades · Automatic detection · Classification · Regression · Multi-objective convolutional learning

2.1 Introduction

Traditionally, many handcrafted features have been successfully used in computer vision tasks, and they often simplify machine learning tasks. Nevertheless, they have a few limitations. These features are often low-level as prior knowledge is hand-encoded and features in one domain do not always generalize to other domains. In recent years, learning feature representations in a supervised manner also known as supervised feature learning is preferred over handcrafted features as they have outperformed the state-of-the-art in many computer vision tasks and have been highly successful. This chapter focuses on feature learning to automatically assess radiographic knee OA severity using convolutional neural networks (CNNs).

Clinically to assess knee OA severity, highly experienced clinicians or radiologists assess the knee joints in X-ray images [1, 2] and assign an ordinal grade based on a radiographic grading scheme. The most commonly used gradings, like the Kellgren and Lawrence (KL) grading scheme and Ahlback system, use distinctive grades (0–4). However, clinical features of knee OA are continuous in nature, and attributing distinctive grades is the subjective opinion of the graders. There are also uncertainties and variations in the subjective gradings. There is a need for automated methods to overcome the limitations arising from this subjectivity, and to improve the reliability in the measurements and classifications [2].

The automatic assessment of knee OA severity has been previously approached in the literature as an image classification problem [1, 3, 4], with the KL grading scale as the ground truth. WNDCHARM,[1] a multi-purpose biomedical image classifier was used to classify knee OA images [4, 5]. High binary classification accuracies (80–91%) have been reported using the WNDCHARM classifier for classifying the extreme stages: grade 0 (normal) versus grade 4 (severe), grade 0 versus grade 3 (moderate). However, the classification accuracies of the images belonging to successive grades are low (55–65%) and the multi-class classification accuracy is low (35%). The overall classification accuracies of knee OA needs improvement for real-world computer aided diagnosis [1, 3, 6].

Radiographic features detected and learned through a computer-aided analysis can be useful to quantify knee OA severity and to predict the future development of knee OA [3]. Instead of manually designing features, the author proposes that learning feature representations using deep learning architectures can be a more effective approach for the classification of knee OA images. Traditionally, hand-crafted features based on pixel statistics, object and edge statistics, texture, histograms, and

[1] Weighted Neighbor Distance using Compound Hierarchy of Algorithms Representing Morphology.

transforms, are typically used for multi-purpose medical image classification [4, 5, 7]. However, these features are not efficient for fine-grained classification such as classifying successive grades of knee OA images. Manually designed or hand-engineered features often simplify machine learning tasks. Nevertheless, they have a few disadvantages. The process of engineering features requires domain related expert knowledge and is often very time consuming [8]. These features are often low-level as prior knowledge is hand-encoded and features in one domain do not always generalize to other domains [9]. The next logical step is to automatically learn effective features for the desired task.

Over the last decade, learning feature representations or feature learning has been preferred to hand-crafted features in many computer vision tasks, particularly for fine-grained classification, because rich appearance and shape features are essential for describing subtle differences between categories [10]. Feature learning refers to techniques that learn to transform raw data input or pixels of an image to an effective representation for further higher-level processing such as object detection, automatic detection, segmentation, and classification. Feature learning approaches provide a natural way to capture cues by using a large number of code words (sparse coding) or neurons (deep networks), while traditional computer vision features, designed for basic-level category recognition, may eliminate many useful cues during feature extraction [10]. Deep learning architectures are multi-layered and they are used to learn feature representations in the hidden layer(s). These representations are subsequently used for classification or regression at the output layer. Feature learning is an integral part of deep learning [8, 11].

Even though many deep learning architectures have been proposed and have existed for decades, in recent times CNNs have become highly successful in the field of computer vision [12, 13]. AlexNet [14] won the ILSVRC[2] in 2012 by a large margin. CNNs have since then become more popular, widely-used and highly-successful in computer vision tasks such as object detection, image recognition, automatic detection and segmentation, content based image retrieval, and video classification [12]. Apart from computer vision tasks, CNNs are finding applications in natural language processing, hyper-spectral image processing, and medical image analysis [12, 15].

CNNs have also recently become successful in many medical applications such as knee cartilage segmentation [16] and brain tumour segmentation [17] in MRI scans, multi-modality iso-intense infant brain image segmentation [18], pancreas segmentation in CT images [19], and neuronal membrane segmentation in electron microscopy images [20]. Inspired by these success stories, the author proposes CNNs for classification of knee OA images and to improve the quantification of knee OA severity and knee OA diagnostic features. The author believes that this can lead to building a real-world knee OA diagnostic system that outperforms the existing approaches.

The remainder of this chapter is organized as follows. Section 2.1.1 introduces knee osteoarthritis (OA), the diagnostic features and the clinical evaluation of knee OA. Section 2.1.2 lists the contributions of this research. Section 2.2 provides an

[2]ImageNet Large Scale Visual Recognition Challenge.

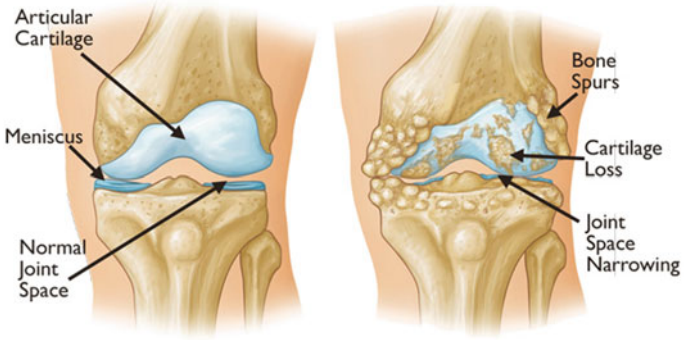


Fig. 2.1 A healthy knee and a knee joint affected with OA

overview of the background, a comprehensive summary of the related work, and a critical analysis of the state-of-the-art in computer aided diagnosis of knee OA. Section 2.3 introduces the public datasets used in this study. Section 2.4 presents the baseline methods using hand-crafted features and the proposed approaches in this chapter for automatic detection of knee joints in the radiographs. Section 2.5 presents the baseline methods and the proposed approaches in this chapter to quantify knee OA severity using CNNs. Section 2.6 concludes this chapter by analyzing the current work and summarizing the research methodology, and providing future directions of research based on the proposed methods.

2.1.1 Knee Osteoarthritis

Knee Osteoarthritis (OA) is a debilitating joint disorder that mainly degrades the knee articular cartilage and in its severe stages it causes excruciating pain and often leads to total joint arthroplasty. In general, knee OA is characterized by joint pain, cartilage wear, and bony growths. Knee OA has a high-incidence among the elderly, obese, and those with a sedentary lifestyle. Early diagnosis is crucial for clinical treatments and pathology [3, 6].

2.1.1.1 Diagnostic Features of Knee OA

Clinically, the major pathological features for knee OA include joint space narrowing, osteophytes formation, and sclerosis [6, 21]. Figure 2.1 shows the anatomy of a healthy knee and a knee affected with osteoarthritis, and the characteristic features of knee OA. The causes for knee OA include mechanical abnormalities such as degradation of articular cartilage, menisci, ligaments, synovial tissue, and sub-chondral bone.

The major clinical features, joint space narrowing and osteophyte formation, are easily visualized using radiographs [6, 7, 22]. Despite the introduction of several imaging methods such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound for augmented OA diagnosis, radiographs have traditionally been preferred [22, 23], and remain as the main accessible tool and "gold standard" for preliminary knee OA diagnosis [1, 3]. Inspired by the previous successful approaches in the literature for early identification [3] and automatic assessment of knee OA severity [4, 6, 23], the focus is on radiographs in this work. More importantly, there are public datasets available that contain radiographs with associated ground truth. Public datasets for knee OA study, such as the OAI[3] and the MOST[4] datasets, provide radiographs with KL scores, and the OARSI[5] readings for distinct knee OA features such as JSN, osteophytes, and sclerosis.

2.1.1.2 Radiographic Classification of Knee OA

Knee OA develops gradually over years and progresses in stages. In general, the severity of knee OA is divided into five stages. The first stage (stage 0) corresponds to normal healthy knee and the final stage (stage 4) corresponds to the most severe condition (see Fig. 2.2). The most commonly used systems for grading knee OA are the International Knee Documentation Committee (IKDC) system, the Ahlback system, and the Kellgren and Lawrence (KL) grading system. The other widely used non-radiographic knee OA assessment system is WOMAC,[6] which measures pain, stiffness, and functional limitation. The public datasets, the OAI and the MOST used in this work, are provided with the KL grades and they are used as the ground truth to classify the knee OA X-ray images.

Kellgren and Lawrence Scores.

The KL grading scale was approved by the World Health Organisation as the reference standard for cross-sectional and longitudinal epidemiologic studies [7, 22, 24, 25]. The KL grading system is still considered the gold standard for initial assessment of knee osteoarthritis severity in radiographs [1, 5–7]. Figure 2.2 shows the KL grading system. The KL grading system categorizes knee OA severity into five grades (grade 0–4). The KL grading scheme for quantifying knee OA severity from X-ray images is defined as follows [1, 5]:

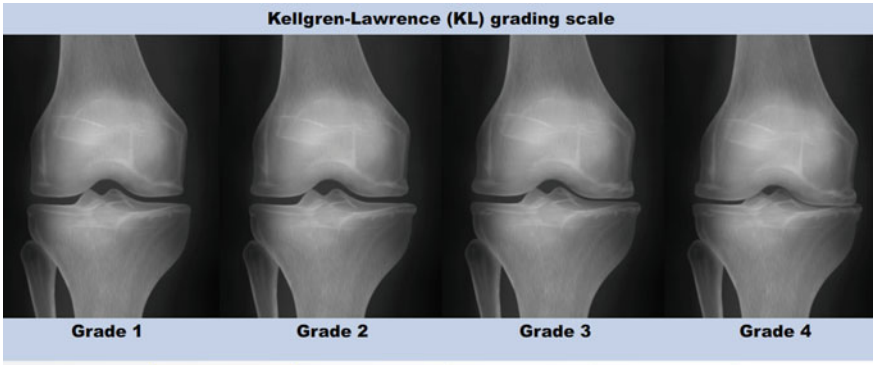
Grade 0: absence of radiographic features (cartilage loss or osteophytes) of OA.
Grade 1: doubtful joint space narrowing (JSN), osteophytes sprouting, bone marrow oedema (BME), and sub-chondral cyst.
Grade 2: visible osteophytes formation and reduction in joint space width on the antero-posterior weight-bearing radiograph with BME and sub-chondral cyst.

[3] Osteoarthritis Initiative.

[4] Multicenter Osteoarthritis Study.

[5] Osteoarthritis Research Society International.

[6] Western Ontario and McMaster Universities Osteoarthritis Index.



CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Fig. 2.2 The Kellgren and Lawrence grading system to assess the severity of knee OA

Grade 3: multiple osteophytes, definite JSN, sclerosis, possible bone deformity.
Grade 4: large osteophytes, marked JSN, severe sclerosis, and definite bone deformity.

2.1.2 Contributions

The research contributions of this work are as follows.

- Proposing a novel and highly accurate technique to automatically detect and localise the knee joints from the X-ray images using a fully convolutional network (FCN).
- Developing a classifier based on a CNN to assess knee OA severity that is highly accurate in comparison to previous methods.
- Proposing a novel approach to train a CNN with a weighted ratio of two loss functions: categorical cross entropy and mean squared error with the natural benefit of predicting knee OA severity in ordinal (0, 1–4) and continuous (0–4) scales.
- Developing an ordinal regression approach using CNNs to automatically quantify knee OA severity in a continuous scale.
- Developing an automatic knee OA diagnostic system i.e. an end-to-end pipeline incorporating the FCN for automatically localising the knee joints and the CNN for automatically quantifying the localised knee joints.

2.2 Related Work and Background

The automatic assessment of knee OA severity from radiographs has been approached as an image classification problem [1, 3, 4]. According to the literature and in the machine learning approach to automatically assess knee OA severity, the first step is to localise the region of interest (ROI) that is to detect and extract the knee joint regions from the radiographs, and the next step is to classify the localised knee joints. First, the different approaches for detecting (or localising) the knee joint regions in the radiographs are outlined. Next, the approaches in the literature to assess knee OA severity are investigated and the focus is on the automated methods. This section concludes with a discussion outlining the limitations in the state-of-the-art methods on automatic detection of knee joints and automatic assessment of knee OA severity, and how these limitations can be addressed.

2.2.1 Detecting Knee Joints in Radiographs

There are several approaches in the literature for detecting and segmenting knee joints and specific parts of the knee such as cartilage, menisci, and bones structures from 3D MRI and CT scan images [26, 27]. Nevertheless, the existing approaches are less accurate for automatically detecting the knee joints in radiographs [27, 28]. According to the literature, detecting knee joints remains a challenging task [27, 29]. In this chapter, automated methods for detecting knee joints in radiographs are investigated. The advantages of automatic methods are discussed and the need to investigate such methods are emphasized. Previous approaches in the literature that investigate the knee joints in radiographs can be categorized into manual, semi-automatic, and fully automatic, based on the level of manual intervention required [26, 27].

2.2.1.1 Manual Methods

Expert radiologists or trained physicians visually examine the knee joint regions and trace the structures using simple image processing and computer vision-based tools in radiographs, and may even use CAD-based measurements for assessing knee OA severity [27]. The expert knowledge-based manual segmentations are useful to build an atlas or template of anatomical structures, which are used to develop advanced interactive and automatic segmentation methods [27]. The knee joints labelled manually are reliable and are often used as ground truth for evaluating automatic methods [30, 31]. Nevertheless, such manual methods are subjective, highly experience-based, and they are laborious and time-consuming when a large number of subjects are to be examined.

There are previous studies in the literature that use manually-defined ROIs (knee joints) in radiographs for assessing knee OA severity. Hirvasniemi et al. [32] quantified the differences in bone density using texture analysis and local binary patterns (LBP) in plain radiographs to assess knee osteoarthritis. Woloszynski et al. [33] developed a signature dissimilarity measure for the classification of trabecular bone texture in knee radiographs. In both these methods, the ROIs are manually marked and extracted for texture analysis.

2.2.1.2 Semi-automatic Methods

Semi-automatic or interactive methods are developed to minimize manual interventions by automating essential steps in the detection and segmentation process [26, 34]. These methods often include manual initialization with low-level image processing, followed by manual evaluations and corrections of the results [35]. The main advantage of the semi-automatic methods are flexibility in manual intervention that allow incorporating expert knowledge, plus the use of advanced computer vision-based tools to automate the essential steps. An expert may improve the detection and segmentation performance through tuning the essential parameters for instance seed region and threshold values in region growing, initial shape of active models, and delineating the required contour [27] to define the region of interest. However, these methods may not be reproducible due to inter-observer or inter-user variations and there is a possibility of oversight or human error in the manual evaluations.

There are some knee OA studies in the literature which use semi-automatic methods to detect the knee joints in radiographs. Knee OA computer aided diagnosis (KOACAD) [6] is an interactive method to measure the joint space narrowing, osteophytes formation and joint angulation in radiographs. In KOACAD, a Roberts filter is used to obtain the rough contour of tibia and femur bone structures and a vertical neighbourhood difference filter is used to identify points with high absolute values of difference of scales. The centre of all the points is calculated and a rectangular region around the centre, of size 480×200 pixels, is selected as the knee joint region. This system has purportedly provided accurate assessment of structural severity of knee OA after detecting the knee joint regions. However, human intervention is required for plotting various lines for the measurement, and automatic detection is not feasible with this system.

Knee images digital analysis (KIDA) is a tool to analyse knee radiographs interactively, proposed by Marijnissen et al. [21]. KIDA quantifies the individual radiographic features of knee OA like medial and lateral joint space width (JSW) measurements, subchondral bone densities and osteophytes. However, this interactive tool can only be used by experts for quantitative measurements and requires expert intervention for objective quantitative evaluation.

Duryea et al. [36] proposed a trainable rule-based algorithm (software) to measure the joint space width between the edges of the femoral condyle and the tibial plateau on knee radiographs. Contours marking the edges of the femur and tibia are

automatically generated. This interactive method can be used to monitor joint space narrowing and the progression of knee osteoarthritis.

2.2.1.3 Automatic Methods

Automatic segmentation methods have become an essential part of computer aided diagnosis and clinical decision support systems [29]. These methods are fast and accurate, and they are highly beneficial in clinical trials and pathology [27]. According to the literature, there have been multiple attempts to automatically localise knee joints in radiographs. Nevertheless, this task still remains a challenge.

Podsiadlo et al. [37] proposed an automated system for the prediction and early diagnosis of knee OA. In this approach, active shape models and morphological operations are used to delineate the cortical bone plates and locate the ROIs in radiographs. This approach is developed for selection of tibial trabecular bone regions in the knee joints as ROIs. Nevertheless, this approach can be extended to localise the entire knee joint. A set of 40 X-ray images are used for training and 132 X-ray images are used for testing in this method. The automatic detections from this method are compared to the gold standard, which contains manually annotated ROIs from the expert radiologists and the similarity indices (SI) are calculated. This method achieved SI of 0.83 for the medial and 0.81 for the lateral regions of the knee joints.

Shamir et al. [1] proposed template matching for automatic knee joint detection in radiographs. Template matching uses predefined joint centre images as templates and calculates Euclidean distances over every patch in an X-ray image using a sliding window. The image patch with the shortest distance is recorded as the detected knee joint centre. After detecting the centre, an image segment of 700×500 pixels around the centre is extracted as the knee joint region. The X-ray images from the BLSA dataset are used in this method. In total 55 X-ray images from each grade are used for the experiments, such that 20 images from each grade for training and 35 images from each grade for testing. Shamir et al. reported that template matching was successful in finding the knee joint centres in all the X-ray images in their dataset.

Anifah et al. [38] investigated template matching and contrast-limited adaptive histogram equalisation for detecting knee joints and quantifying joint space area. In total 98 X-ray images are used in this method. The detection accuracy achieved by this method varies from 83.3 to 100% for the left knees and 60.4 to 100% for the right knees. Template matching is a simple and relatively fast method. However, this method is ad hoc, entirely based on the set of templates used and is unlikely to generalise well for larger datasets.

Recently, Tuilpin et al. [29] investigated a SVM-based method to automatically localise knee joints in plain radiographs. This method uses knee anatomy-based region proposals, and the best candidate region from the proposals are selected using histogram of oriented Gradients (HOG) as feature descriptors and a SVM. This method generalises well in comparison to the previous methods and shows reasonable improvement in automatic detections with mean intersection over union (IOU) of 0.84, 0.79 and 0.78 on the public datasets MOST, Jyvaskyla, and OKOA.

2.2.2 Assessing Radiographic Knee OA Severity

The key pathological features of knee OA include joint space narrowing, osteophytes (bone spurs) formation, and sclerosis (bone hardening) [6, 21]. All these features are implicitly integrated in composite scoring systems, like Kellgren and Lawrence (KL) grading system, to quantify knee OA severity [6, 21], and the OARSI readings provide the gradings of distinct knee OA features. There are two common approaches for assessing knee OA severity in plain radiographs: (1) quantifying the distinct pathological features of knee OA, and (2) automatic classification based on composite scoring systems such as KL grades.

2.2.2.1 Quantitative Analysis

The most conventional system to assess radiographic knee OA severity has been KL gradings [3, 5, 6]. Nevertheless, some researchers [6, 21] argue that categorical systems like KL gradings are limited by incorrect assumptions that the progression of distinct OA features like JSN and osteophytes formation is linear and constant, their relationships are proportional, and such grading systems are less sensitive to small changes in distinct features. Therefore, quantification of individual features of knee OA is required to overcome the problems with KL gradings and to improve the overall radiographic assessment of knee OA [6, 21]. The OARSI has published a radiographic atlas of individual features to assess and to quantitatively evaluate the knee OA features [6].

Interactive methods (KOACAD [6] and KIDA [21]) measure individual knee OA radiographic features such as joint space width (JSW), osteophyte area, sub-chondral bone density, joint angle, and tibial eminence height as continuous variables. These measurements were compared to KL gradings and significant differences were found between healthy knees and knees with OA. In this context, a trainable rule-based algorithm has also been proposed [36] to measure the minimum joint space width (mJSW) between the edges of the femoral condyle and the tibial plateau, and thus to monitor the progression of knee OA. Podsiadlo et al. [37] have used a slightly different approach for quantitative knee OA analysis. In this method, the trabecular bone regions of the tibia are automatically located as the ROI after delineating the cortical bone plates using active shape models, followed by fractal analysis of bone textures for the diagnosis of knee OA. In a similar approach, Lee et al. [39] use active shape models to detect the tibia and femur joint boundaries, and calculate anatomical geometric parameters to diagnose knee OA.

Even though these methods are simple to implement, objective, and purpotedly accurate in evaluating radiographic knee OA, a great deal of manual intervention is required. Hence, these methods become very time-consuming and laborious when large numbers of subjects are to be investigated. Furthermore, the measurements from these methods are prone to inter- and intra-observer variability.

2.2.2.2 Automatic Classification

After the introduction of radiography-based semi quantitative scoring systems like KL gradings, the assessment of radiographic knee OA severity has been approached as an image classification problem [2, 3, 40–42]. According to the literature, the most common approach to classify knee OA images includes two steps: (1) extracting image features from the knee joints, and (2) applying a classification algorithm on the extracted features. A brief review of such approaches follows.

Subramoniam et al. [40, 41] investigated two methods using: (1) the histograms of local binary pattern extracted from knee images and a k-Nearest neighbour classifier [40] and (2) Haralick features extracted from the ROI of knee images and a SVM [41]. Thomson et al. [2] proposed an automated method that uses features derived from tibia and femur bone shapes, and image textures extracted from the tibia with a simple weighted sum of the outputs of two random forest classifiers. Deokar et al. [42] investigated an artificial neural network based approach for knee OA images classification using grey level co-occurrence matrix (GLCM) textures, shape, and statistical features. Even though these methods claim high accuracy, the datasets are not publicly available and these datasets contain only a few hundred radiographs. The classification accuracies of all these methods for public datasets like the OAI and the MOST need to be studied to derive conclusive results.

In this context, there are two approaches that use large public datasets like the OAI: (1) WNDCHRM, and (2) an artificial neural network-based scoring system. Shamir et al. proposed WNDCHRM, a multi purpose medical image classifier to automatically assess knee OA severity in radiographs [1, 3]. A set of features based on polynomial decompositions, high contrast, pixel statistics, and textures are used in WNDCHRM. Besides extracting features from raw image pixels, features extracted from image transforms like Chebyshev, Chevbyshev-Fourier, Radon, and Gabor wavelets are included to expand the feature space [1, 4, 5]. From the entire feature space, highly informative features are selected by assigning feature weights based on a Fisher discriminant score for all the extracted features [1, 3, 5]. WNDCHRM uses a variant of the k-Nearest Neighbour classifier.

Yoo et al. [43] have built a self-assessment scoring system and an artificial neural network (ANN) model for radiographic and symptomatic knee OA risk prediction. In a recent approach, Tiulpin et al. [44] presented a new computer-aided diagnostic approach based on deep Siamese CNNs, which were originally designed to learn a similarity metric between pairs of images. However, rather than comparing image pairs, the authors extend this idea to similarity in knee x-ray images (with 2 symmetric knee joints). Splitting the images at the central position and feeding both knee joints into a separate CNN branch allows the network to learn identical weights for both branches. They outperform the previous approaches by achieving an average multi-class testing accuracy score of 66.7% on the entire OAI dataset.

2.2.3 Discussion

According to the literature, the automatic quantification of knee OA severity involves two steps: (1) automatically detecting the ROI, and (2) classifying the detected knee joints. Many previous studies investigated automatic methods for both localisation and classification of knee joint images, but still these tasks remain a challenge.

The common approaches in the literature for automatic detection of knee joints in radiographs include template matching [1, 38], active shape models and morphological operations [37], and a classifier-based sliding window method [29]. Template matching and active shape models based approaches do not generalise well and are slow for large datasets. Classifier-based methods that use hand-crafted features are subjective and the classification accuracy is influenced by the choice of extracted features. Therefore, there is still a need for an automated method for detecting knee joints in radiographs which gives high accuracy and precision. A deep learning based method for this is investigated in this chapter.

There are several approaches in the literature for knee OA image classification that have extracted and tested many image features, such as Haralick textures [41], Gabor textures [2], GLCM textures [42], local binary patterns [40], shape, and statistical features of knee joints [42]. There is even an approach that uses a large set of features based on pixel statistics, object and edge statistics, texture, histograms, and transforms [4, 5, 7]. Different classifiers have been tested for knee OA images classification such as k-Nearest Neighbour [1, 40], SVM [41], and random forest classifiers [2]. However, all these approaches have achieved low multi-class classification accuracy, and in particular classifying successive grade knee OA images still remains a challenging task. There is a need for a highly accurate real world automated system that can be used as a support system by clinicians and medical practitioners for knee OA diagnosis.

In recent years, many methods using manually designed or hand-crafted features have been outperformed by approaches that learn feature representations using deep neural networks. In particular, convolutional neural networks (CNN) have become highly successful in many computer vision tasks like object detection, face recognition, content based image retrieval, pose estimation, and shape recognition, and even in medical applications such as knee cartilage segmentation in MRI scans [16], brain tumour segmentation in magnetic resonance imaging (MRI) scans [17], multi-modality iso-intense infant brain image segmentation [18], pancreas segmentation in CT images [19], and neuronal membrane segmentation in electron microscopy images [20]. CNNs for automatically quantifying knee OA severity is investigated in this work. The next section introduces the public knee OA datasets.

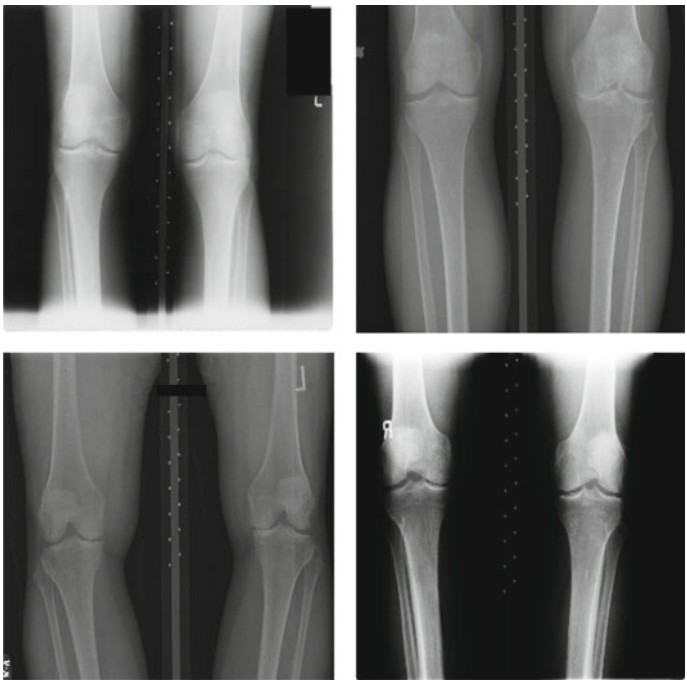


Fig. 2.3 Samples of bilateral PA fixed flexion knee OA radiographs

2.3 Public Knee OA Datasets

The data used for the experiments and analysis in this study are bilateral PA fixed flexion knee X-ray images. Figure 2.3 shows some samples of knee X-ray images from the dataset. Due to variations in X-ray imaging protocols, there are some visible artefacts in the X-ray images (Fig. 2.3).

The datasets are from the Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST) in the University of California, San Francisco. These are standard public datasets used in knee osteoarthritis studies.

2.3.1 OAI Dataset

The baseline cohort of the OAI dataset contains MRI and X-ray images of 4,746 participants. In total 4,446 X-ray images are selected from the entire cohort based on the availability of KL grades for both knees as per the assessments by Boston University X-ray reading centre (BU). In total there are 8,892 knee images. Figure 2.4 shows the distribution as per the KL grades.

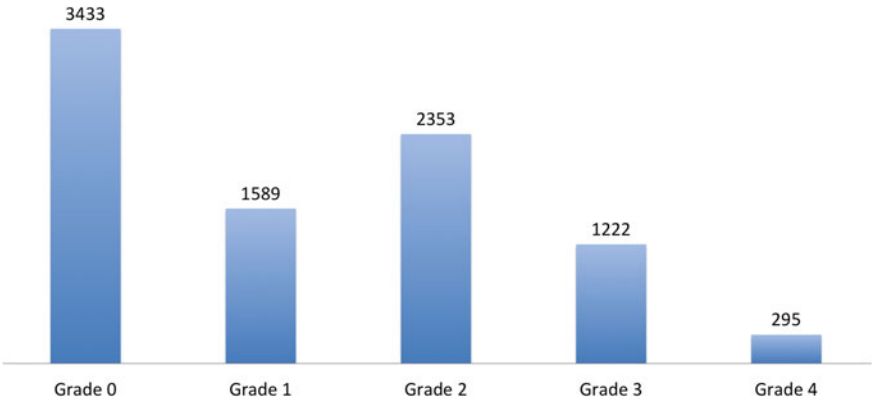


Fig. 2.4 The OAI baseline data set distribution based on KL grades

2.3.2 MOST Dataset

The MOST dataset includes lateral knee radiograph assessments of 3,026 participants. In total 2,920 radiographs are selected in this study based on the availability of KL grades for both knees as per baseline to 84-month longitudinal knee radiograph assessments. There are 5,840 knee images in this dataset. Figure 2.5 shows the distribution as per KL grades.

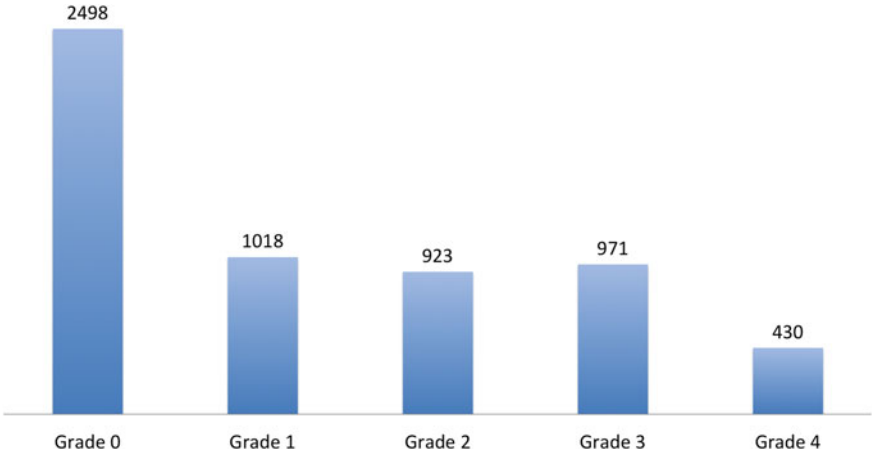


Fig. 2.5 The MOST data set distribution based on KL grades

2.4 Automatic Detection of Knee Joints

Classification of knee OA images and the assessment of severity conditions can be achieved by examining the characteristic features of knee OA: variations in the joint space width and the osteophytes (bone spurs) formations in the knee joints [6]. Radiologists and medical practitioners examine only the knee joint regions in the X-ray images to assess knee OA. Hence, the region of interest (ROI) for classifying knee OA images is only the knee joint regions (left and right knees). Figure 2.6 shows the ROI in a X-ray image. The author believes that it is better to focus on the ROI instead of the entire X-ray image for accurate classification and this is also computationally economical. For these reasons, automatically detecting and extracting the knee joint regions from the X-ray images becomes an essential pre-processing step, before classification.

2.4.1 Baseline Methods

First, template matching for automatic detection of the knee joints [3] is implemented as a baseline. Next, the author proposes an SVM-based automatic detector for this. The implementation details and outcomes of these methods are discussed in this section.

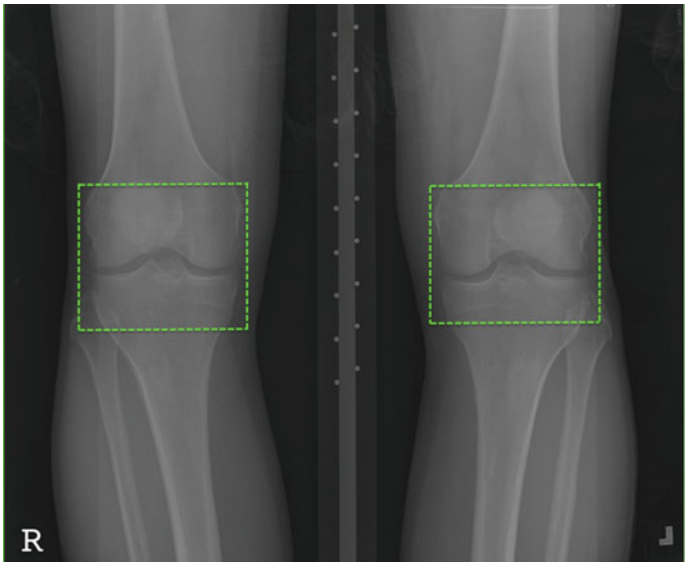


Fig. 2.6 A knee OA X-ray image with the region of interest: the knee joints

2.4.1.1 Template Matching

In digital image processing, template matching is a technique for finding portions of an image that are similar to a standard template image. Shamir et al. [3] proposed this approach for automatically detecting the centre of the knee joints. As a baseline, the template matching approach is adapted. The steps involved in this method are as follows:

- First, the radiographs are downscaled to 10% of the original size and subjected to histogram equalisation for intensity normalisation. This step is followed as proposed by Shamir et al. [3].
- An image patch (20 × 20 pixels) containing the centre of the knee joint is taken as a template. 5 image patches are taken from each grade, so that in total 25 patches are pre-selected as templates. Figure 2.7 shows the pre-selected knee joint centres of size 20 × 20 pixels extracted from the knee joint images as templates.
- Each image is scanned by an overlapping (20 × 20) sliding window. For each location at an interval of 10 pixels, distances (Euclidean) between an image patch (20 × 20 pixels) and 25 pre-selected templates (patches with knee joint centre) are computed using;

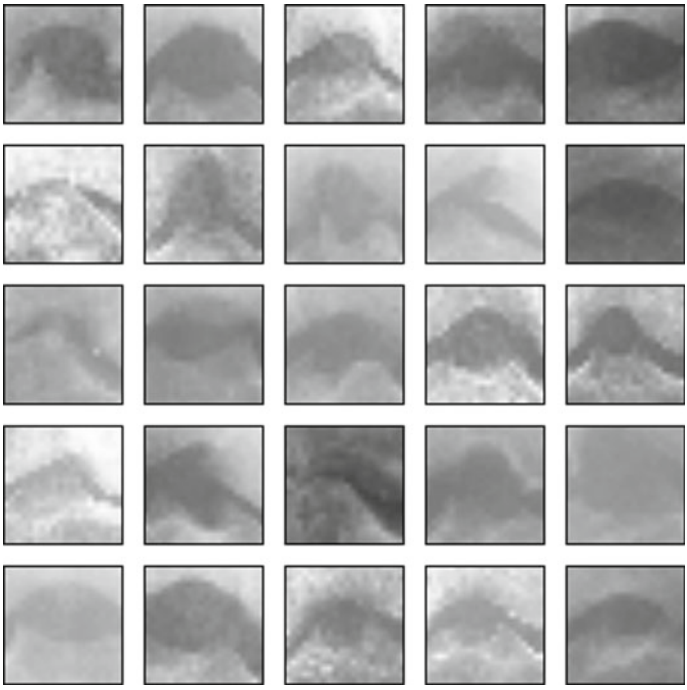


Fig. 2.7 Pre-selected knee joint centres (20 × 20 pixels) extracted from knee joint images for template matching

$$dist_{i,w} = \sqrt{\sum_{y=1}^{20}\sum_{x=1}^{20}(I_{x,y} - W_{x,y})^2},$$

where $I_{x,y}$ is the intensity of pixel (x, y) in the knee joint image I, $W_{x,y}$ is the intensity of pixel (x, y) in the sliding window, and $dist_{i,w}$ is the Euclidean distance between the knee joint image (I) and the sliding window W.

- In total, 25 different distances are calculated at each location of the sliding window for the 25 templates, and the shortest among the 25 distances is recorded.
- The window with the smallest Euclidean distance is selected as the centre of the knee joint after scanning the image with a sliding window, and a fixed size region (700 × 500 pixels) around this centre is extracted as the knee joint region from the X-ray image.
- The input X-ray images are horizontally split in half to isolate left and right knees separately and the sliding window is run on both halves.

Experiments and Results. For the experiments on template matching, the baseline data sample of 200 progression and incidence cohort subjects under the knee OA study is used. This dataset contains in total 191 X-ray images (382 knee joints) and it is a subset of the large OAI dataset.

In this implementation, five different sets of templates (each set with 25 templates) are used to show the influence of templates on knee joint detections. The templates are selected from a separate training set. Visual inspection is used to evaluate the results of template matching by plotting a bounding box (20 × 20 pixels) on the image patch that recorded the shortest Euclidean distance after template matching. Table 2.1 shows the total number of true positives (the detected knee joint centres), the total number of false positives and the precision.

It is clearly evident from the results (Table 2.1) that template matching is not precise in detecting the knee joints and that the detections are heavily dependent on the choice of templates. The number of templates was increased to 50, but still there was no further improvement in the results. The reason for low-performance of the template matching is that the computations are mainly based on the intensity level difference between an image patch and a template. There are also possibilities for image patches not around the knee joint, having the shortest Euclidean distance to

Table 2.1 Detection of knee joint centres using template matching method

Templates	True positives	False positives	Precision (%)
Set 1	87	295	22.8
Set 2	78	304	20.4
Set 3	99	283	25.9
Set 4	**116**	**266**	**30.3**
Set 5	55	327	14.4

a template in the set and thus, being detected as matches. In the next section, a new SVM-based method is investigated to improve the detection of the knee joints.

2.4.1.2 SVM-Based Detection

Standard template matching is not scalable and produces poor detection accuracy on large datasets like the OAI. We proposed a classifier-based model to automatically detect the knee joints in the X-ray images [45]. The idea is to use well-known Sobel edge detection [46] for detecting the knee joints. The two major steps involved in this method are (1) training a classifier and (2) developing a sliding window detector.

Training a Classifier. First, image patches (20 × 20 pixels) are generated from the input X-ray images. The image patches containing the knee joint centre (20 × 20 pixels) are used as positive samples and randomly sampled patches excluding the knee joint centre are used as negative samples. In total, 200 positive and 600 negative samples are used. The image patches (samples) are split into training (70%) and test (30%) sets. Sobel horizontal image gradients are extracted as features from all these samples to train a classifier. The powerful and well-known SVM is used for classification. A linear SVM is fitted with default parameters (C = 1, and linear kernel), using Sobel horizontal image gradients as the features.

Before settling on Sobel horizontal image gradients as features, the state-of-the-art features such as histogram of oriented gradients, Tamura and Haralick textures, and the Gabor features were tested. The HOG features are highly accurate and efficient in object detection and human detection [47]. The Tamura, Haralick, and Gabor features are highly influential and top-ranked among the features used in WNDCHRM for knee OA image classification [1, 3, 5, 6]. The Sobel operator or Sobel filter uses vertical and horizontal image gradients to emphasise the edges in images [46]. From these, the horizontal image gradients are used as the features for detecting the knee joints centres. Intuitively, the knee joint images primarily contain horizontal edges that are easy to detect.

Sliding Window Detector. To detect the knee joint centre from both left and right knees, input images are split in half to isolate left and right knees separately. A sliding window (20 × 20 pixels) is used on either half of the image, and the Sobel horizontal gradient features are extracted for every image patch. The image patch with the maximum score based on the SVM decision function is recorded as the detected knee joint centre, and the area (200 × 300 pixels) around the knee joint centre is extracted from the input images using the corresponding recorded coordinates. Figure 2.8 shows an instance of a detected knee joint and the extracted ROI in a X-ray image.

Results and Discussion. In total, 200 image patches with the knee joint centres as positive samples and 600 image patches that exclude the centre of knee joint as negative samples are used. These images are split into training (70%) and test (30%) sets. Fitting a linear SVM with the training data produced a 5-fold cross validation accuracy of **95.2**% and an accuracy of **94.2**% for the test data. Table 2.2 shows

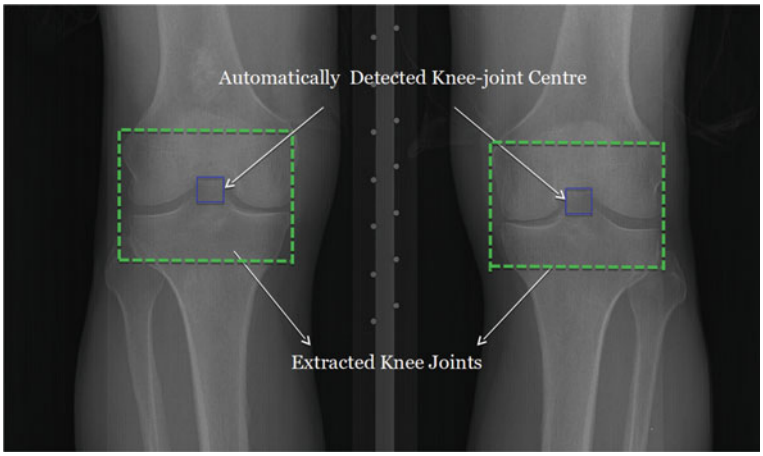


Fig. 2.8 Detecting the knee joint centres and extracting the knee joints

Table 2.2 Classification metrics of the SVM for detection

Class	Precision	Recall	F_1 score
Positive	0.93	0.84	0.88
Negative	0.95	0.98	0.96
Mean	0.94	0.94	0.94

Table 2.3 Comparison of template matching and the proposed SVM-based method

Method	$JI = 1$ (%)	$JI \geq 0.5$ (%)	$JI > 0$ (%)
Template matching	0.3	8.3	54.4
Proposed method	1.1	38.6	**81.8**

the precision, recall, and F_1 scores of this classification. To evaluate the automatic detection, the ground truth is generated by manually annotating the knee joint centres (20×20 pixels) in 4,446 radiographs using an annotation tool that we developed, which recorded the bounding box (20×20 pixels) coordinates of each annotation.

The well-known Jaccard index (JI) is used to give a matching score for each detected instance. The Jaccard index JI(A,D) is given by,

$$JI(A, D) = \frac{A \cap D}{A \cup D} \tag{2.1}$$

where A, is the manually annotated and D is the automatically detected knee joint centre using the proposed method.

Table 2.3 shows the resulting average detection accuracies based on thresholding of Jaccard indices. The mean JI for the template matching and the classifier methods

are **0.1** and **0.36**. From Table 2.3, it is evident that the proposed method is more accurate than template matching. This is due to the fact that template matching relies upon the intensity level difference across an input image. Thus, it is prone to matching a patch with small Euclidean distance that does not actually correspond to the knee joint centre. Also, the templates are varied in a set, and it is observed that the detection is highly dependent on the choice of templates. Template matching is similar to a k-nearest neighbour classifier with $k = 1$.

The reason for higher accuracy in the proposed method is the use of horizontal edge detection instead of intensity level differences. The knee joints primarily contain horizontal edges and thus are easily detected by the classifier using horizontal image gradients as features. The proposed method is approximately 80× faster than template matching; for detecting all the knee joints in the dataset comprising 4, 446 radiographs, the proposed method took ∼9 min and the template matching method took ∼798 m.

Despite sizeable improvements in accuracy and speed using the proposed approach, detection accuracy still falls short. Therefore, manual annotations for the incorrect detections from this method were substituted to investigate KL grade classification performance independently of knee joint detection. The next Section describes the proposed methods for automatically localising the knee joint region using fully convolutional neural networks.

2.4.2 Fully Convolutional Network Based Detection

A typical CNN architecture consists of three main types of layers: convolutional, pooling and fully-connected or dense layers. A fully convolutional network (FCN) is similar to a CNN, but the fully-connected layers are replaced by convolutional layers [48]. A FCN consists of mostly convolutional layers and if pooling layers are used, then suitable up-sampling layers are added before the last convolutional layer. The two major differences of FCNs over CNNs can be summarised as:

– FCNs are trained end-to-end to make pixel-wise predictions [48]. Even the decision-making layers at the last stage of the network use learned convolutional filters.
– The input image size need not be fixed as there are no fully-connected layers in the FCN. CNNs with fully connected layers can operate only on a fixed size input.

FCNs have achieved great success in semantic segmentations of general images [48]. Recent approaches using FCNs for medical image segmentation show promising results [49–51]. Motivated by this, the use of FCN is investigated in this chapter for automatically detecting the knee joints. Two approaches are developed for localising the knee joints: (1) training a FCN to detect the centre of knee joints and extract a fixed-size region around the detected centre, and (2) training a FCN to detect the ROI and thus extract the knee joints directly.

2.4.2.1 Localisation with Reference to Knee Joint Centre

In the initial approach to localise the knee joints in X-ray images using a FCN, a similar strategy to template matching and the SVM based methods is followed; that is to detect the centre of knee joints and to extract the ROI with reference to the detected centres. Figure 2.9 shows the steps involved in this method: training a FCN to detect the knee joint centres (20×20 pixels), computing the coordinates of the centres from the FCN output, and extracting a fixed size region as knee joints. In the next section, the experimental data and the ground truth used to train the FCNs are introduced.

Dataset and Ground Truth Generation. The data used for the experiments are taken from the baseline cohort of the OAI dataset. In total 4,446 X-ray images are selected from the entire dataset based on the availability of KL grades for both knee joints. The knee joint centres in all these X-ray images are manually annotated, after downscaling to 10% of the actual size. Binary masks of size 20×20 pixels are marked around the knee joint centres using the annotations. Figure 2.10 shows an instance of an input X-ray image and the binary mask annotations corresponding to the knee joint centres. The image patches from the masked region i.e. the knee joint centres, are taken as positive training samples and the patches from rest of the image are taken as the negative training samples to train an FCN. The dataset is split into training (3,333 images) and test (1,113 images) sets.

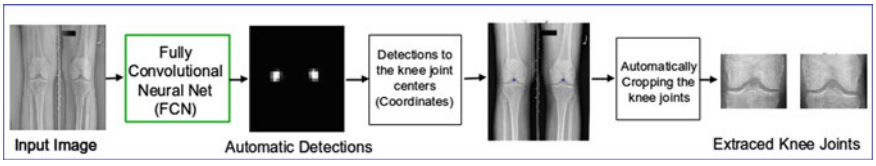


Fig. 2.9 Automatic localisation of knee joints with reference to the centre of the knee joints

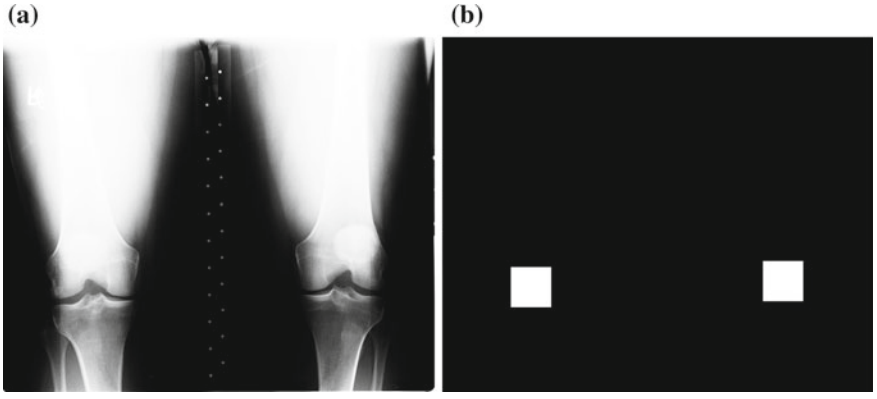


Fig. 2.10 **a** An input X-ray image and **b** the binary mask annotations for knee joint centres

Training Fully Convolutional Neural Networks. To start, a FCN is configured with a lightweight architecture containing 4 convolutional layers followed by a fully convolutional layer, which is a convolutional layer with a kernel size [1 × 1] and that uses a *sigmoid activation*. FCNs use fully convolutional layers at the last stage to make pixel-wise predictions [48]. Table 2.4 shows the network configuration in detail. Each convolution layer is followed by a ReLU layer.

The network parameters are trained from scratch with training samples of knee OA radiographs from the OAI dataset. The dataset is split into training (3,333 images) and test (1,113 images) sets. The ground truth for training the network are binary images with masks specifying the ROI: the knee joints. The network is trained to minimise the total *binary cross entropy* between the predicted pixels and the ground truth. *Stochastic gradient descent* (SGD) with default parameters: learning rate = 0.01, decay = $1e^{-6}$, momentum = 0.9, and nesterov = True, is used. The network is trained for 40 epochs and the batch size is 10. Figure 2.11 shows an instance of the test input, the ground truth and the output (pixel-wise predictions) of the FCN. From the predictions of this FCN, it is observed that the network is able to slightly detect the edges of the knee joints and these are promising initial results. In an attempt to improve the detections, the FCN configurations are experimented and for this the hyper-parameters of the network are tuned.

Receptive Field. When dealing with high-dimensional inputs such as images, it is impractical to connect neurons in the current level to all the neurons in the previous volume. Instead, each neuron is only connected to a local region of the input volume.

Table 2.4 Initial FCN configuration for detecting the knee joint centres

Layer	Kernel	Kernel size
conv1	32	3 × 3
conv2	32	3 × 3
conv3	64	3 × 3
conv4	64	3 × 3
conv5	1	1 × 1

Fig. 2.11 An instance of input, ground truth and output (predictions) of FCN

Table 2.5 FCN for detecting the knee joint centres

Layer	Kernel	Kernel size
conv1	32	7×7
conv2	64	3×3
conv3	96	3×3
conv4 (fullyConv)	1	1×1

The spatial extent of this connectivity is a hyper-parameter called the receptive field of the neuron [52]. The receptive field size, otherwise termed the effective aperture size of a CNN, shows how much a convolutional node sees of the input pixels (patch) that affects a node's output. The effective aperture size depends on kernel size and strides of the previous layers. For instance, a 3×3 kernel can see a 3×3 patch of the previous layer and a stride of 2 doubles what all succeeding layers can see.

The receptive field size of neurons in the final layer of the FCNs is calculated and used to analyse the output of FCNs and the overall detection results. The receptive field size of a neuron in the final layer (conv5) of the initial FCN configuration (Table 2.4) is 9, which is low and may be a reason for poor performance of this network. Larger convolutional kernel sizes to increase the receptive field of the network is investigated. The forthcoming Section will show that a network (Table 2.9) with larger receptive field gives the best results for detecting the knee joint centres.

Tuning the FCN Hyper-parameters. VGG-M-128 [53], the deep convolutional neural network developed by the Oxford visual geometry group (VGG) uses kernel size 7×7 in the first convolutional layer and 5×5 in the following convolutional layer. Inspired by this, kernel sizes of 5×5, and 7×7 for the first convolutional layer are tested retaining the other settings. The kernel size 7×7 gives better results in this configuration. This is because of the larger receptive field size of the 7×7 kernel in comparison to the 3×3 kernel.

Next, the experiments are conducted by varying the number of convolutional layers and also the number of filters (kernel) in a convolutional layer, before obtaining the configuration that gave the best results based on visual observations. Table 2.5 shows the configuration of the network derived from the initial configuration and the receptive field size of a neuron in the final layer (conv4) is 11. The networks are trained with 3,333 images and tested on 1,113 images from the OAI dataset.

There is an improvement in the detections using this network in comparison to the previously tested configurations. Figure 2.12 shows an instance of the output predictions of this network. To quantitatively evaluate the automatic detections, the well-known Jaccard Index is used.

Quantitative Evaluation. A simple contour detection is used and the Jaccard index i.e. the overlap statistics calculated by the Intersection over Union (IoU) to evaluate the automatic detections of the FCN. The steps involved are as follows:

Fig. 2.12 An input image, ground truth, and outcome of the final FCN

- First, the objects are detected i.e. the knee joint regions from the output image of the FCN using simple contour detection [54]. Contours can be explained simply as a curve joining all the continuous points (along the boundary), having the same colour or intensity. The contours are a useful tool for shape analysis and simple object detection and recognition. In this method, first the images are converted to binary by applying Otsu's threshold. Next, the contours of the objects or shapes in the binary image are automatically detected and recorded [54].
- Next, the detected objects in the image are sorted based on the area and from these the top two are selected. This is to eliminate noise or other faint edges picked up by the FCN.
- The centroids of the largest two detected regions are recorded as the knee joint centres.
- A binary mask of 20 × 20 pixels size is marked around each detected knee joint centre.
- The Jaccard index is computed for each image with the masks of predicted centres and the masks predefined using manual annotation i.e. the labels used for training FCN.

In total 1,113 X-ray images (2,226 knee joints) are included in the test set. The FCN with the final configuration detects 1,851 knee joints in the test set with Jaccard index $\geq$0.5, the accuracy of detection is **83.2**% with a mean 0.66 and standard deviation 0.18. This is an improvement in comparison to previous approaches but still falls short of perfect detections. The pooling and up-sampling layers in the FCN are varied and experimented in an attempt to improve the detection accuracy. This may help to increase the receptive field size and in turn improve the overall detections.

FCN with Pooling and Up-sampling Layers. Two max pooling layers with stride 2 and up-sampling by a factor of 4 are included to the previous configuration (Table 2.5). Table 2.6 shows the FCN architecture in detail. Each convolutional layer is followed by a ReLU activation.

Figure 2.13 shows the output of this network for a test image. On visual observation, the output image contains less noise and the detections are improving compared to the previous approaches, even though the output image resolution is low. This is

Table 2.6 FCN with pooling and up-sampling layers

Layer	Kernel	Kernel size	Strides
conv1	32	7×7	1
maxPool2	–	2×2	2
conv3	64	3×3	1
maxPool4	–	2×2	2
conv5	96	3×3	1
upSamp6	–	4×4	1
conv7 (fullyConv)	1	1×1	1

Fig. 2.13 Prediction of the FCN with max pooling and up-sampling layers

Table 2.7 FCN with 3 convolution-pooling stages for detecting the knee joint centres

Layer	Kernel	Kernel size	Strides
conv1	32	7×7	1
maxPool2	–	2×2	2
conv3	32	3×3	1
maxPool4	–	2×2	2
conv5	64	3×3	1
maxPool6	–	2×2	2
conv7	96	3×3	1
upSamp8	–	8×8	1
conv9 (fullyConv)	–	1×1	1

due to the inclusion of pooling and up-sampling stages to the network and this has increased the receptive field size of the final layer (conv7) to 34. The number of convolutional-pooling stages is increased, to see if there is improvement in the detections. Table 2.7 shows the architecture of this network in detail.

From the output of this FCN, it can be observed that the detections become more precise in comparison to the previous networks even though the resolution is low in comparison to the previous networks. Figure 2.14 shows an instance of the input test image, ground truth and the FCN output.

Fig. 2.14 Predictions of the FCN with 3 convolution-pooling stages

Table 2.8 Detection accuracy of FCN based on Jaccard Index

Jaccard index	JI $\geq$ 0.25 (%)	JI $\geq$ 0.5 (%)	JI $\geq$ 0.75 (%)
Detection accuracy	98.5	96.7	39.6

The outcomes of this FCN are evaluated using Jaccard index and the detection accuracy is **96.7**%, that is in total 2,152 out of 2,226 knee joints are detected with a Jaccard index $\geq$ 0.5. The Jaccard index mean is 0.74 and standard deviation is 0.13. The detection accuracy is high in comparison to the previous networks. Table 2.8 shows the detection accuracy of the FCN for the Jaccard index values at 0.25, 0.5 and 0.75.

This FCN (Table 2.7) has three convolutional-pooling stages. A configuration with 4 convolutional-pooling stages followed was tested by adding an up-sampling layer with kernel size (16 $\times$ 16). There was no improvement in the detection accuracy for this configuration.

Best Performing FCN for Detecting the Knee Joint Centres. Before settling on the final architecture, experiments were done by varying the number of convolution stages, the number of filters and kernel sizes in each convolution layer. The best performing FCN (Table 2.9) was selected based on a high detection accuracy on the test data. This network was trained with the OAI dataset containing 4,444 knee radiographs. The dataset was split into a training set containing 3,333 knee images and test set containing 1,113 knee images. The validation set (10%) was taken from the training set. The effective aperture size of this FCN (Table 2.9) for a node in the last convolutional layer (before up-sampling) is 66. The aperture size for the previous networks shown in Table 2.7 is 42 and Table 2.6 is 34. For the other tested configurations the effective aperture size is even lower (less than 30).

Table 2.9 shows the configuration of the best performing FCN for detecting the knee joint centres. This FCN is based on a lightweight architecture and the network parameters (in total 214,177) are trained from scratch. The network consists of 4 stages of convolutions with a max-pooling layer after each convolutional stage, and the final stage of convolutions is followed by an up-sampling and a fully-convolutional layer. The network uses a uniform [3 $\times$ 3] convolution and [2 $\times$ 2]

Table 2.9 Best performing FCN for detecting the knee joint centres

Layer	Kernel	Kernel size	Strides
conv1	32	3×3	1
maxPool1	–	2×2	2
conv2-1	32	3×3	1
conv2-2	32	3×3	1
maxPool2	–	2×2	2
conv3-1	64	3×3	1
conv3-2	64	3×3	1
maxPool3	–	2×2	2
conv4-1	96	3×3	1
conv4-1	96	3×3	1
upSamp5	–	8×8	1
conv5 (fullyConv)	–	1×1	1

Table 2.10 Detection accuracy of the best performing FCN

Jaccard index	JI $\geq$ 0.25 (%)	JI $\geq$ 0.5 (%)	JI $\geq$ 0.75 (%)
Detection accuracy	98.9	97.1	43.3

max pooling. Each convolution layer is followed by a ReLU activation layer. After the final convolution layer, an $[8 \times 8]$ up-sampling is performed as the network uses 3 stages of $[2 \times 2]$ max pooling. The up-sampling is essential for an end-to-end learning by back propagation from the pixel-wise loss and to obtain pixel-dense outputs [48]; when pooling layer(s) and strides more than one are used in the network. The final layer is a fully convolutional layer with a kernel size of $[1 \times 1]$ and uses a sigmoid activation for pixel-based classification. The input to the network is of size $[256 \times 256]$.

This network was trained to minimise the total binary cross entropy between the predicted pixels and the ground truth using *stochastic gradient descent* (SGD) with default parameters: learning rate = 0.01, decay = $1e^{-6}$, momentum = 0.9, and nesterov = True. This network was trained for 40 epochs with a batch size of 32. The validation (10%) data was taken from the training set. Figure 2.15 shows the learning curves when training this network and decrease in the validation and training losses.

Table 2.10 shows the results of the best performing FCN. This network achieved a detection accuracy of **97.1**%, in total 2,162 knee joints out of the 2,226 test samples detected with a Jaccard index 0.5. The Jaccard index mean is 0.76 and standard deviation is 0.12.

Error Analysis. The results of the best performing FCN (Table 2.10) show 99% detection accuracy for a Jaccard index $\geq$ 0.1, in total 2,205 out of 2,226 knee joints

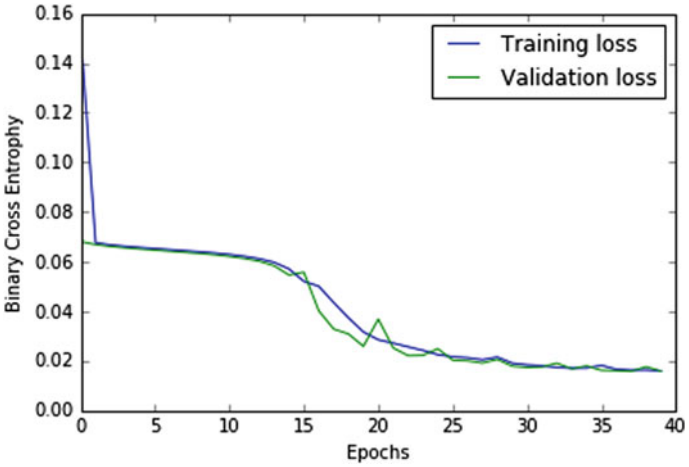


Fig. 2.15 Training and validation losses of the FCN

are successfully detected. On observing the failed detections: 1% (in total 21 knee joints), there are two patterns.

1. The output of the FCN is very faint or no detections at all. Figure 2.16 shows two instances of input X-ray images, masks defining the knee joint centres as ground truth, and output of the best performing FCN with faint detections. The input images with variations in the local contrast and local luminance due to the imaging protocol variations appear to be the main cause for this error. Histogram equalisation is used as a pre-processing step to adjust the contrast of the input images. Even though this adjusts the contrast globally in an image, there are still contrast variations in portions of the image. Local contrast enhancement algorithms [55] or adaptive histogram equalisation [56] can be used to normalise the images for variations in the local contrast and local luminance.
2. The FCN output picks up noise along with the knee joints. Figure 2.17 shows two instances of input X-ray images, masks defining the knee joint centres as ground truth, and output of the best performing FCN with noise. The reason for this error appears to be due to the variations in the imaging protocol and resolution of the X-ray images, and presence of artefacts in the input X-ray images. Intuitively, the FCN uses horizontal edge detection along with other features to detect the knee joints. The artefacts with predominant horizontal edges are picked up by the FCN along with the centre of knee joints. When simple contour detection is applied on the FCN output, instead of the knee joints the artefacts are also detected.

Automatically Extracting the Knee Joints. After training FCNs to automatically detect the centre of the knee joints, the next step is to extract the ROI i.e. the knee joints with reference to the detected centres. The initial goal is to train an end-to-end network for localising the knee joints i.e. to directly predict the bounding box co-

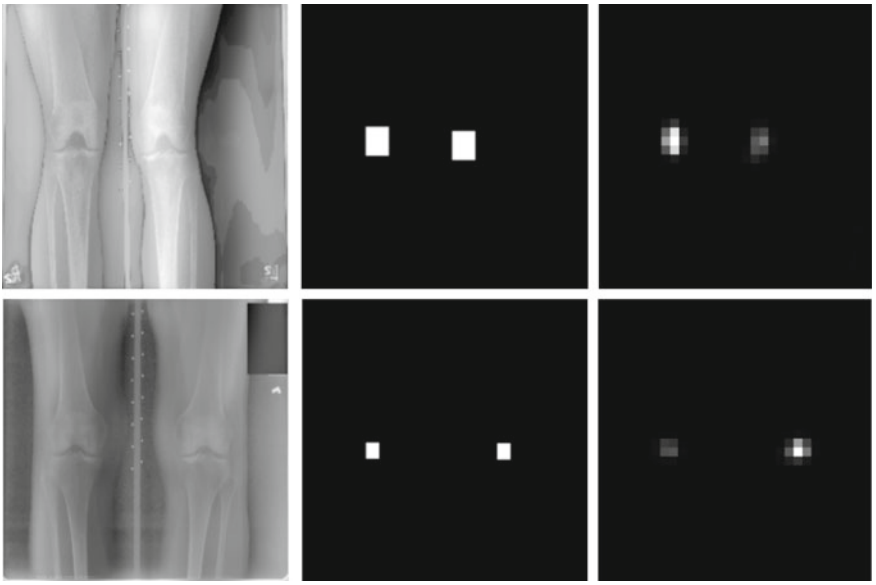


Fig. 2.16 Error analysis: X-ray images, ground truth, FCN output—weak detections

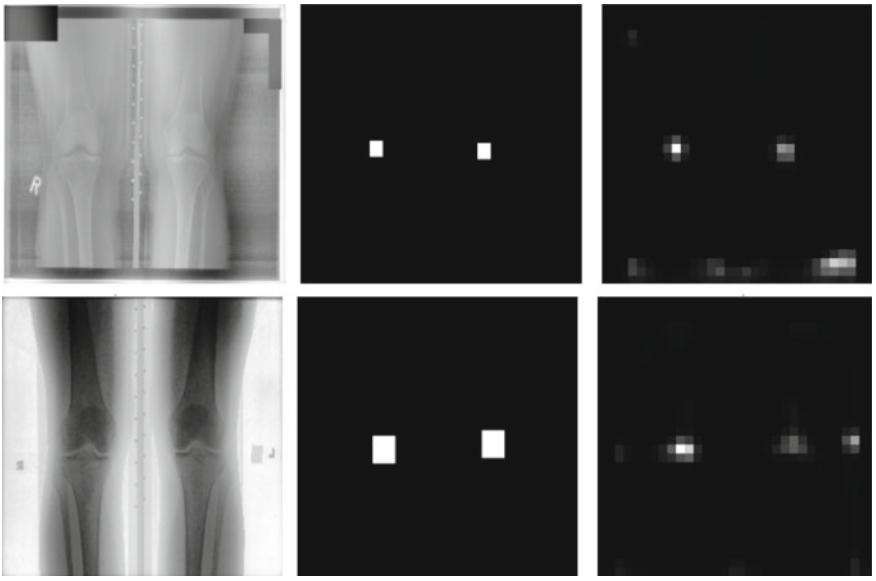


Fig. 2.17 Error analysis: X-ray images, ground truth, FCN output—detections with noise

ordinates of the knee joints from the input X-ray images. A bounding box regression is investigated [57] that is a network trained on top of the FCN (Table 2.9) output, to achieve this. First, CNNs are trained with the masks (20×20) of knee joint centres as the input (256×256) and the bounding box coordinates of the left knee joint (x_1, y_1) and right knee joint (x_2, y_2) as the ground truth (labels). Next, CNNs are trained with the X-ray images as input and the targets (labels) are the bounding box coordinates instead of the binary masks. However in both the experiments, the networks trained to predict the bounding boxes give low accuracy. On considering the overall knee joint centres, there is no large variations in the centre coordinates. The reason for the low accuracy is that the networks are not learning discernible features to predict the bounding box coordinates. This affects the overall performance of the localisation. Therefore, a simple approach based on contour detection is used to calculate the centres and extract the knee joints. Figure 2.18 shows an X-ray image with the centres, the left and the right knee joints extracted from the X-ray image using the centroids. The steps involved in this method are as follows.

- First, the contour detection [54] is used on the FCN output to calculate the spatial coordinates of the knee joint centres. In the contour detection method, first the input images (FCN output) are converted to binary by applying Otsu's threshold. Next, the contours from the binary image are automatically detected and recorded. Finally, the centroids are calculated from the detected knee joint regions.
- The knee OA radiographs are resized to 2560×2560, that is 10 times the size of the FCN output 256×256.
- The detected knee joint centres are up-scaled to a factor of 10.
- Fixed size regions (640×560) are extracted around the up-scaled centres as the knee joint regions. After testing and visualising different sizes for the knee joint crop, image patch with the size (640×560) is found to be mostly suitable and containing the required ROI for further quantification. Figure 2.18 shows an instance of the extracted left and right knee joints.

Localisation Results. The results of the FCN are compared to the previous methods: template matching and SVM-based method to automatically detect the centre of the knee joints. All these methods are evaluated based on the Jaccard index (JI). Table 2.11 shows the detection accuracy of the knee joint centres using FCN, SVM-based method, and template matching. The results show that the proposed method using FCN clearly outperforms the previous methods. This also demonstrates that feature learning using an FCN is a better approach for detecting the knee joints than using hand-crafted features such as Sobel gradients and the template matching method that is sensitive to intensity level variations. However, the extracted knee joints from this method have some limitations.

Limitations of this Method. In all three approaches; FCN-based, SVM-based and template matching, the centre of the knee joints are detected and these are used as reference for automatically localising the knee joints. There are some limitations in extracting a fixed size region as the ROI with reference to the detected centres due to the variations in the resolution of the X-ray images and the variations in the size of the knee joints.

Fig. 2.18 A knee X-ray image with the detected centres and the extracted left and right knees

Table 2.11 Comparison of methods used for localising the centre of the knee joints

Method	JI > 0 (%)	JI ≥ 0.5 (%)	JI ≥ 0.75 (%)	Mean	Std. dev.
Template matching	54.4	8.3	3.1	0.1	0.2
SVM-based method	81.8	38.6	10.2	0.36	0.31
Fully ConvNet	**98.9**	**97.1**	**43.3**	**0.76**	**0.12**

All the images are resized to a fixed size $2,560 \times 2,560$ and extract a fixed size region 640×560 around the detected centres as the ROI. Due to this scaling issue, portions of the knee joints are omitted in the automatic extraction of the ROI. Figure 2.19 shows such instances. Figure 2.20 shows the corresponding actual ROIs.

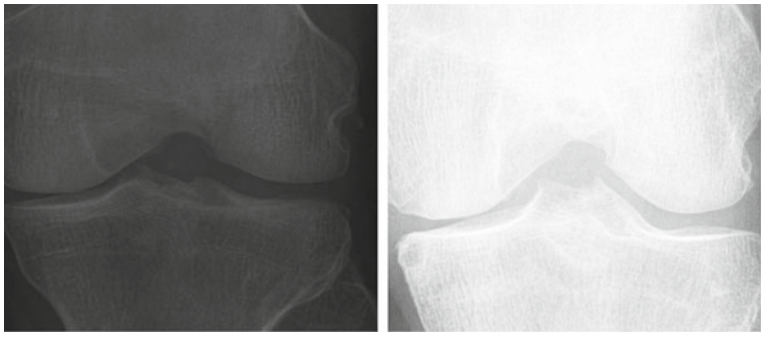


Fig. 2.19 Anomalies in the automatic extraction of the ROI

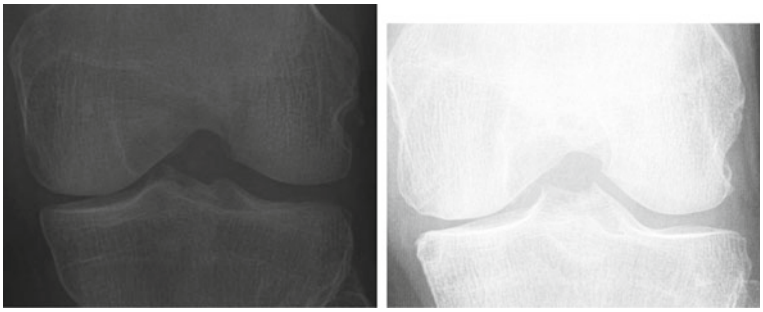


Fig. 2.20 The actual ROI for the knee joints in Fig. 2.19

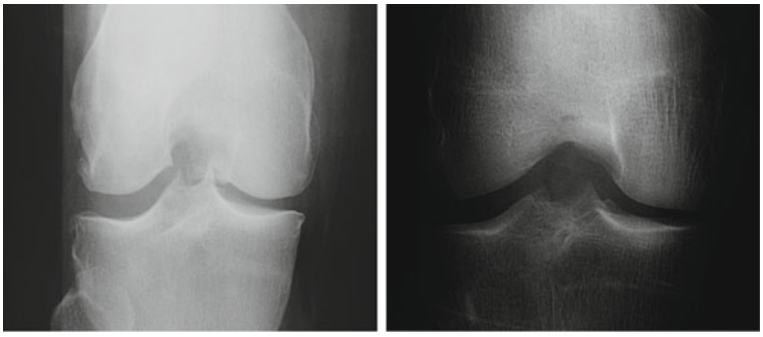


Fig. 2.21 Variations in the aspect ratio of the extracted knee joints

Due to the varying sizes of the knee joints and a fixed size region being extracted as the ROI, there are differences in the aspect ratio of the extracted and the actual ROI. Figure 2.21 shows instances where the knee joints are small in comparison to the fixed size region extracted as the ROI. Figure 2.22 shows the actual ROIs.

The classification of the automatically extracted knee joints is compared to the manually extracted knee joints. There is a decrease in the accuracy by a margin of 3–4% when using the automatically extracted knee joints with reference to the detected centres. The discrepancies in the localisation of knee joints affects the overall classification of the knee OA images. To overcome these limitations, as the next approach FCNs are trained to detect the ROI itself, instead of detecting the knee joint centres.

2.4.2.2 Localising the Region of Interest

The previous methods to localise the knee joints in the X-ray images with reference to the automatically detected centres have certain limitations. To overcome these

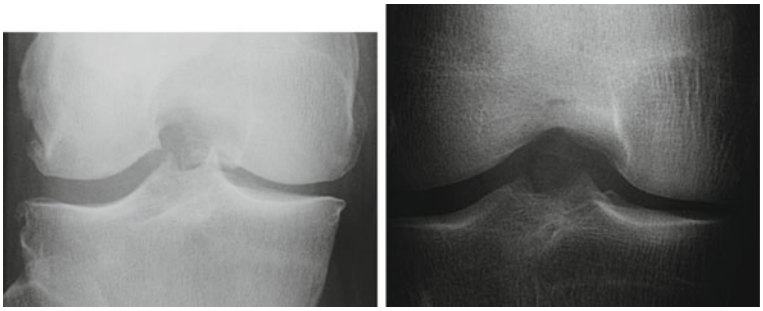


Fig. 2.22 The actual ROI for the extracted knee joints in Fig. 2.21

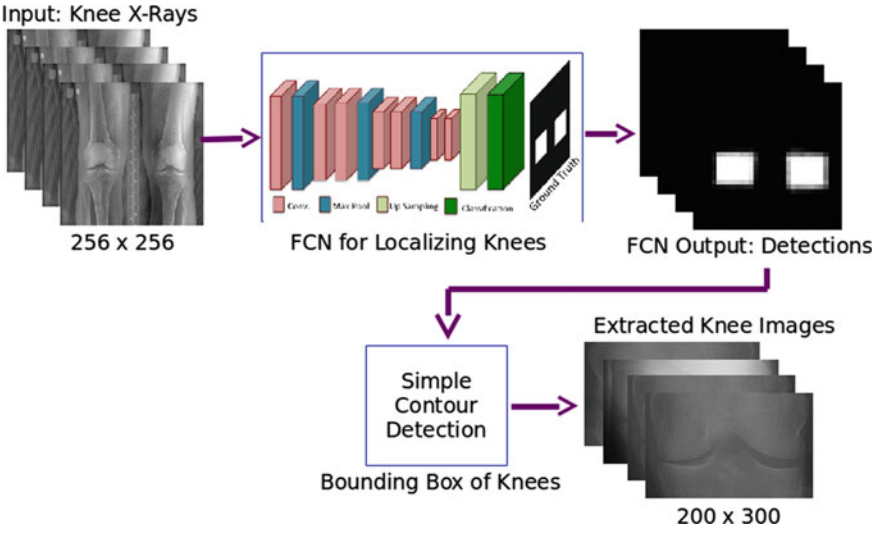


Fig. 2.23 Automatic localisation of the region of interest

limitations and to improve the localisation, FCNs are trained to detect the ROI directly [58]. Figure 2.23 shows the steps involved in this method.

Dataset and Ground Truth. For the experiments in this approach, a new dataset from the MOST is used along with the data from the previous experiments, the baseline cohort of the OAI dataset. In total 4,446 X-ray images are selected from the OAI dataset and 2,920 X-ray images from the MOST dataset based on the availability of KL grades for both knee joints. The full ROI is manually annotated in all these X-ray images, after downscaling to 10% of the actual size. The down-sampling of the images is necessary to reduce the computational costs. Binary masks are generated based on the manual annotations. Figure 2.24 shows an instance of an input X-ray image and the binary mask annotations corresponding to the ROI. The image patches from the masked region (the knee joints) are taken as positive training samples and

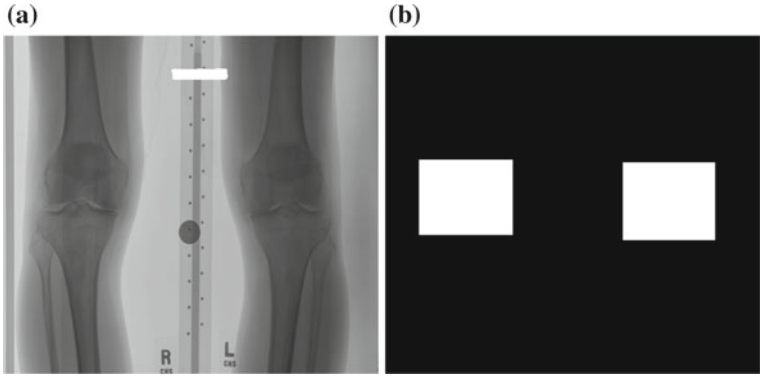


Fig. 2.24 **a** An input X-ray image and **b** the binary mask annotations for the region of interest

the patches from rest of the image are taken as the negative training samples to train a FCN. The datasets are split into a training/validation set (70%) and test set (30%). The training and test samples from the OAI dataset are 3,146 images and 1,300 images, and from the MOST dataset are 2,020 images and 900 images.

Training the FCN. First, a FCN is trained using the same architecture (Table 2.9) from the previous approach to detect the ROI. Initially, the network is trained with training samples from OAI dataset and test it with OAI and MOST datasets separately. Next, the training samples are increased by including the MOST training set where the test set is a combination of both OAI and MOST test sets. This network is trained to minimise the total binary cross entropy between the predicted pixels and the ground truth using the adaptive moment estimation (Adam) optimiser with default parameters: initial learning rate $(\alpha) = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$. Adam optimiser gives faster convergence than standard SGD. Figure 2.25 shows the learning curves converging to small loss when training this network. Figure 2.26 shows the output of this network for a test image.

A few other network configurations are tested by varying the number of convolutional-pooling stages, convolutional layers in each stage and the number of convolutional kernels in a convolutional layer. There was no further improvement in the detection accuracy on the validation set. Therefore, this configuration was settled as the final network for localising the knee joints.

Quantitative Evaluation. The Jaccard index, i.e. the intersection over Union (IoU) of the automatically detected and the annotated knee joint is used to quantitatively evaluate the automatic detections. For this evaluation, all the knee joints in both the OAI and MOST datasets are manually annotated using a fast annotation tool. Table 2.12 shows the number (percentage) of knee joint correctly detected based on the Jaccard index (JI) values greater than 0.25, 0.5 and 0.75 along with the mean and the standard deviation of JI. Table 2.12 also shows detection rates on the OAI and MOST test sets separately.

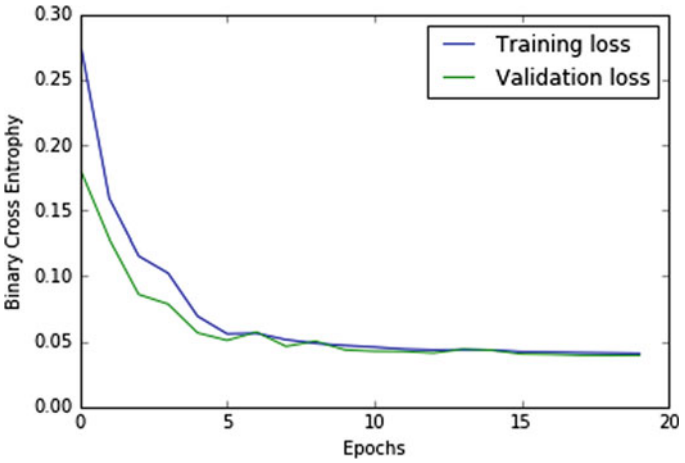


Fig. 2.25 Training and validation losses of the FCN

Fig. 2.26 An input X-ray image, ground truth and output prediction of the FCN

Table 2.12 Comparison of automatic detection based on the Jaccard index (JI)

Test data	JI > 0 (%)	JI ≥ 0.5 (%)	JI ≥ 0.75 (%)	Mean	Std. Dev.
OAI	100	100	88	0.82	0.06
MOST	99.7	98.8	80.6	0.80	0.09
Combined OAI-MOST	**100**	**100**	**92.2**	**0.83**	**0.06**

Considering the anatomical variations of the knee joints and the imaging protocol variations, the automatic detection with a FCN is highly accurate with 100% detection accuracy for JI ≥ 0.5 and 92.2% (4,056 out of 4,400) of the knee joints for J ≥ 0.75 being correctly detected.

Qualitative Evaluation. Figures 2.27, 2.28, and 2.29 show a few instances of successful knee joint detections with the JI values for the left and right knee detections. Detecting the ROI directly gives high accuracy (100%) in comparison to the previous

Fig. 2.27 Qualitative evaluation: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.98, right knee with JI = 0.888

Fig. 2.28 Qualitative evaluation: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.879, right knee with JI = 0.969

Fig. 2.29 Qualitative evaluation: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.768, right knee with JI = 0.984

method (Sect. 2.4.2) to detect the knee joint centres and extracting a fixed size region as the ROI. The FCN in this method learns features from a relatively larger region (the actual ROI) in comparison to the previous method where the FCN is confined to learn features from a small region (20×20), the centre of the knee joints, and therefore, the detections are more accurate.

Fig. 2.30 Error analysis: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.83, right knee with JI = 0.398. The implants in the right knee is the reason for this localisation error

Fig. 2.31 Error analysis: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.473, right knee with JI = 0.837. The implants in the left knee is the reason for this localisation error

Error Analysis. This method is highly accurate with 100% detection accuracy for a JI $\geq$ 0.5. Nevertheless, there are a few anomalies in the FCN detections due to variations in the imaging protocols, presence of artefacts and noise in the input images. Figures 2.30 and 2.31 show two instances where one knee has undergone joint-arthoplasty and the knee implants are visible in the X-ray images, and due to this the FCN detections are distorted. Figures 2.32, 2.33 and 2.34 show a few instances of X-ray images with noise and presence of artefacts due to imaging protocols. This adversely affects the FCN detections.

Extracting the Knee Joints. The bounding boxes of the knee joints are calculated using simple contour detection from the output predictions of the FCN. After converting the FCN output to binary image using Otsu's threshold, the contours are detected using simple image analysis by calculating the zero order moments [54], which gives the perimeter of the detected object. The contours are recorded as bounding boxes. The knee joints are extracted from knee OA radiographs using the bounding boxes. The bounding boxes are up-scaled from the output of the FCN that is of size [256 $\times$ 256] to the original size of each knee OA radiograph, before extracting the knee joints so that the aspect ratio of the knee joints is preserved.

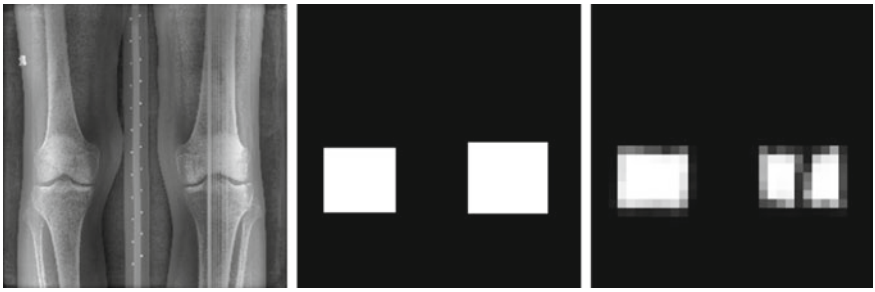


Fig. 2.32 Error analysis: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.887, right knee with JI = 0.356. The noise in the right knee causes this localisation error

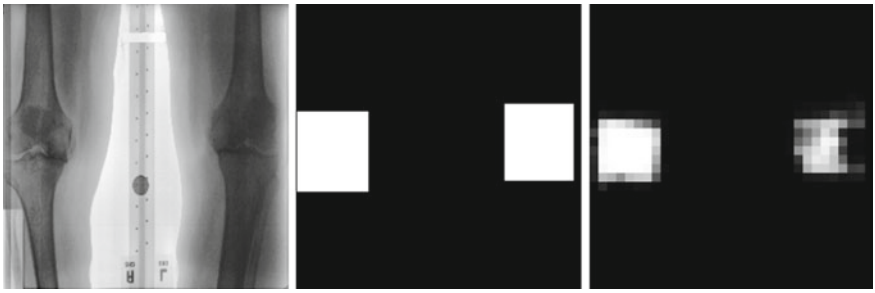


Fig. 2.33 Error analysis: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.681, right knee with JI = 0.488. The localisation error in this image is due to the variation in the imaging protocol

Fig. 2.34 Error analysis: an input X-ray image, ground truth, and FCN detections: left knee with JI = 0.768, right knee with JI = 0.507. The variations in the local contrast and luminance affects the localisations

2.4.3 Summary and Discussion

Automatically localising the knee joints in X-ray images is an important and an essential step before quantifying knee OA severity. Previously, template matching was implemented as a baseline method to localise the knee joints, proposed by Shamir et al. [1, 3], and it was shown that the detection accuracy is low ($\sim 30\%$) in this method for large datasets like OAI. To improve the localisation, a SVM-based method with Sobel horizontal image gradients as features was proposed in this Section. This method showed a large improvement in detection accuracy (82%) but still falls short of perfect localisation. The anomalies in localised knee joints can affect the step involving classification of the localised knee joints to quantify knee OA severity.

Instead of using hand-crafted features, a deep learning-based solution was proposed in this Section to further improve localisation. FCNs were trained to automatically detect and extract the knee joints. All three methods: template matching, SVM-based and FCN-based were evaluated using a common metric: the Jaccard Index. This method achieved almost perfect detection with 100% accuracy for a Jaccard Index 0.5 and an accuracy of 92% for a Jaccard index greater than equal to 0.75. The author believes this performance is sufficient to localise and extract the knee images for classification. As such further improvements are left as future work. The localisation performance may be improved by including additional pre-processing steps to remove the artefacts and noise in the images, and to normalise the local contrast variations in the images. Using additional data for learning and data augmentation may improve the localisation performance.

2.5 Automatic Assessment of Knee OA Severity

Previous work on automated assessment of knee OA severity approached it as an image classification problem [2, 3, 40–42]. Previous methods have tested many hand-crafted features based on pixel statistics, textures, edge and object statistics, and transforms [3, 4, 6, 7, 40, 41]. Many classifiers such as the SVM [41], the k-nearest neighbour classifier [40], the weighted neighbour nearest classifier [3, 4], the random forest classifiers [2], and even artificial neural networks (ANN) [42, 43] have been tested for knee image classification. As a baseline (in Sect. 2.4.1), the state-of-the-art features successful in computer vision tasks, such as histogram of oriented gradients [47], local binary patterns [59], and Sobel Gradients [46] are tested. These features are not included in the previous studies to assess knee OA severity. All the previous approaches based on hand-crafted features give low multi-class classification accuracy when classifying knee images, and in particular classifying fine-grained successive knee OA grades remains a challenge. As a baseline, the state-of-the-art CNNs features (in Sect. 2.5.2.2) are also tested for knee images classification on a

small baseline data set from OAI and this approach gave promising results. Motivated by this, the use of CNNs are investigated for quantifying knee OA severity.

2.5.1 Baseline for Classifying Knee OA Radiographs

2.5.1.1 WNDCHRM Classification

WNDCHRM is an open source utility for biological image analysis and medical image classification [3, 4, 60]. In WNDCHRM, a generic set of image features based on pixel statistics (multi–scale histograms, first four moments), textures (Haralick and Tamura features), factors from polynomial decomposition (Zernike polynomials), and transforms (Radon, Chebyshev statistics, Chebyshev-Fourier statistics) are extracted. For feature selection, every feature is assigned a *Fisher score*[7] and 85% of the features with lowest *Fisher scores* are rejected and the remaining 15% of the features are used for classification [3].

Experiments. The dataset used for the initial experiments to classify knee OA images using WNDCHRM are taken from the baseline data sample of 200 progression and incidence cohort. After histogram equalisation and mean normalisation of the X-ray images, the knee joints are extracted manually from the radiographs. The extracted knee joints are split into training (70%) and test (30%) sets. The WNDCHRM command line program is used to classify the extracted knee joint images. WNDCHRM uses a variant of k-nearest neighbour classifier.

Results and Discussion. The baseline dataset is not balanced and there are only 44 samples available in KL grade 4. Given the limited number of images in this class, only a small number of images are used for training and testing (35 images for training and 9 images for testing) for multi-class classification. For other classifications 100 images are used for training and 30 images for testing.

It is evident from the results (Table 2.13) that the multi-class classification accuracy and successive grades classification accuracies are very low. The reason for low classification accuracy is that the features used for classification are not capable of capturing the minute structural and morphological variations in the knee joints between the successive grades. Next, the state-of-the-art hand-crafted features are investigated in an attempt to improve the classification accuracy.

2.5.1.2 Classification Using Hand-Crafted Features

Histogram of oriented gradients (HOG), local binary patterns, and Sobel Gradients are tested for classifying knee OA images [1, 3, 5]. These features are not used in

[7]Fisher score is one of the widely used methods for determining the most relevant features for classification.

Table 2.13 Results of WNDCHRM classification

Classification	Grades	Accuracy (%)
Binary	G0 versus G1	66.7
	G1 versus G2	48.3
	G2 versus G3	60
	G3 versus G4	55
	G0 versus G2	48.3
	G0 versus G3	70
Multi-class	G0 to G4	28.3
	G0 to G3	35.8

the previous studies. HOG describes the local object shape and appearance within an image by the distribution of intensity gradients or edge directions and the HOG descriptor was successful in human detection [47]. LBP is powerful for image texture classification. LBP uses local spatial patterns and grey scale contrast as measures for texture classification [59].

Experiments with HOG, LBP and Sobel descriptors. Once again the images from the baseline data sample of 200 progression and incidence cohort is used. The HOG, LBP and Sobel descriptors are extracted from the knee joint images and a SVM is used for classification. Table 2.14 shows the classification results of successive grades of knee OA images using a SVM and the feature space included the HOG, LBP and Sobel gradients.

There is no large improvement in the classification accuracies using HOG, LBP and Sobel gradients features with SVM classification from the previous results with the WNDCHRM classification. To improve the classification, the features space is expanded by including highly effective and top-ranked features from the WND-CHRM classification.

Expanding the feature space. The features based on pixel statistics and textures such as Tamura, Haralick, Gabor and Zernike are used for classification. These features are used in the WNDCHRM classification. Tamura texture features represent

Table 2.14 Classification results of the proposed methods using hand-crafted features

Grades	WNDCHRM (%)	SVM classification with hand-crafted features			
		HOG (%)	LBP (%)	Sobel (%)	Combining all (%)
G0 versus G1	**66.7**	53.3	58.3	58.3	55
G1 versus G2	48.3	48.3	53.3	**58.3**	51.6
G2 versus G3	60	60	60	56.7	**63.3**
G3 versus G4	55	**65**	**65**	50	**65**

contrast, coarseness and directionality of an image [61]. Haralick features are the statistics computed on the co-occurrence matrix of an image [62]. Gabor textures are based on Gabor wavelets and the image descriptors are computed using Gabor transform of an image [63]. Zernike features are obtained by the Zernike polynomial approximation of an image [64]. The feature space for classification is formed by simple concatenation of all the extracted features into a super vector following the early fusion approach.

Results and Discussion. First, a SVM is used with the extracted features for classifying knee OA images. Next, a k-nearest neighbour classifier and support vector regression (SVR) are tested for classification. In total, 100 knee joint images are taken for training and 30 for test set in each grade. Table 2.15 shows the classification accuracy of the WNDCHRM classifier and the classification using kNN, SVM, and SVR.

When comparing the classification results of the proposed methods (SVM, kNN, and SVR) to the WNDCHRM classification, for some cases the results are slightly better and promising. Nevertheless, there is a need for a more significant improvement in the classification results. In these experiments, a subset of features from WNDCHRM, such as Tamura and Haralick texture features, Gabor wavelet features, and Zernike features, were extracted and used for classification. In addition to these features HOG, LBP, and Sobel Gradients were tested. It was found that by further expanding the feature space by including features from WNDCHRM based on transforms such as Radon, Chebyshev, FFT, and Wavelet, and compound image transforms such as Chebyshev-FFT, Chebyshev-Wavelet, and Wavelet-FFT classification can be improved. However, the author believes that learning feature representations can be more effective for fine-grained knee OA classification. In the following section, the state-of-the-art CNN features are investigated for classifying knee OA images.

Table 2.15 Classification results of WNDCHRM and the proposed methods using hand-crafted features

Grades	WNDCHRM (%)	Proposed methods		
		kNN (%)	SVM (%)	SVR (%)
G0 versus G1	**66.7**	55	60	60
G1 versus G2	48.3	**61.7**	46.7	48.3
G2 versus G3	**60**	51.7	55	**60**
G3 versus G4	**55**	35	50	45
G0 versus G2	48.3	46.7	55	**56.7**
G0 versus G3	**70**	48.3	58.3	60

2.5.2 *Automatic Quantification Using Convolutional Neural Networks*

First, the use of off-the-shelf CNNs are investigated for quantifying knee OA severity through classification and regression. Two approaches are followed for this: (1) using a pre-trained CNN for fixed feature extraction, and (2) fine-tuning pre-trained CNN following a transfer learning approach. WNDCHRM, an open source utility for medical image classification [3, 4, 60] is used for benchmarking the classification results obtained from the proposed methods.

Next, three new methods are investigated to automatically quantify knee OA: (1) training a CNN from scratch for multi-class classification of knee OA images; (2) training a CNN to optimise a weighted ratio of two loss functions categorical cross-entropy for multi-class classification and mean-squared error for regression; and (3) training a CNN for ordinal regression of knee OA images. The results from these methods are compared to the previous methods. The classification results using both manual and automatic localisation of knee joints are also compared.

2.5.2.1 Off-the-shelf CNNs

The use of well-known off-the-shelf CNNs such as the VGG-16 network [53], and comparatively simpler networks like VGG-M-128 network [65], and BVLC reference CaffeNet [66, 67] (which is very similar to the widely-used *AlexNet* model [14]) are investigated to classify knee OA images. These networks are pre-trained for general image classification using a very large dataset: the ImageNet LSVRC dataset [68] which contains more than 1.2 million images in 1000 classes. Initially, features are extracted from the convolutional, pooling, and fully-connected layers of VGG-16, VGG-M-128, and BVLC CaffeNet, and used to train linear SVMs to classify knee OA images.

The pre-trained networks are fine-tuned for knee OA images classification motivated by the transfer learning approach [69]. Transfer learning is adopted as the OAI dataset is small, containing only a few thousand images. In transfer learning, a base network is first trained on external data, and then the weights of the initial n layers are transferred to a target network [69]. The new layers of the target network are randomly initialised following the Xavier weight initialisation procedure [70]. The random weights initialisations increase the likelihood of the training algorithms during the backpropagation to obtain a global solution through the gradient descent instead of settling to a nearest local solution.

Intuitively, the lower layers of the networks contain more generic features such as edge or texture detectors useful for multiple tasks, whilst the upper layers progressively focus on more task specific cues [67, 69]. This approach is used for both classification and regression, adding new fully-connected layers, and backpropagation is used to fine-tune the weights for the complete network on the target loss.

2.5.2.2 Classification Using CNN Feature Extraction

The VGG-16 network [53] is trained with the OAI dataset. Features are extracted from different layers of the VGG net such as fully-connected (fc7), pooling (pool5), and convolutional (conv5-2) layers to identify the most discriminating set of features. Linear SVMs (LIBLINEAR [71]) are trained with the extracted CNN features for classifying knee OA images, where the ground truth are images labelled with KL grades. Next, the use of simple pre-trained CNNs such as VGG-M-128 [65] and the BVLC CaffeNet model [66] are investigated for classifying the knee OA images. These networks have fewer layers and parameters in comparison to the VGG-16 network. The features are extracted from the fully-connected, pooling, and convolutional layers, using the VGG-M-128 net and the BVLC reference CaffeNet.

Experiments and Results. The knee joint images are split into training (∼70%) and test (∼30%) set based on the distribution of each KL grade. Features are extracted from fully-connected, pooling, and convolution layers of VGG-16, VGG-M-128, and BVLC CaffeNet. Linear SVMs are trained individually for binary and multi-class classifications on the extracted features. WNDCHRM is used for benchmarking the classification results from the proposed methods in this chapter [3, 4, 60]. WNDCHRM is trained with the same training data so that the classification results from WNDCHRM and CNN features can be compared. The knee OA images are classified in three ways as follows. Classifying healthy knee images (grade 0) with the progressive stages (grade 1, 2, 3, and 4), classifying the images belonging to the successive stages (grade 0 vs. 1, grade 1 vs. 2, …) and multi-class classification to classify all the stages of knee OA images.

Table 2.16 shows the test set classification accuracies achieved by WNDCHRM and the CNN features. The CNN features consistently outperform WNDCHRM for classifying healthy knee samples against the progressive stages of knee OA. The features from conv4 layer with dimension $512 \times 13 \times 13$ and pool5 layer $256 \times 13 \times 13$ of VGG-M-128 net, and conv5 layer with dimension $512 \times 6 \times 6$ and pool5 layer with dimension $256 \times 6 \times 6$ of BVLC reference CaffeNet give higher classification accuracy in comparison to the fully-connected fc6 and fc7 layers of VGG nets and CaffeNet. Intuitively, the lower layers capture more discriminative low-level features such as edge or shape detectors, and the higher layers tend to contain high-level features specific to object classes as per the training data. Features are also extracted from lower layers such as pool4, conv4-2, pool3, pool2 and train classifiers on top of these features. As the dimension of the bottom layers are high, the training time is increased, however, no improvement in classification accuracy is observed.

In a fine-grained classification task such as knee OA image classification, the accuracy of classifying successive classes tends to be low, as the variations in the progressive stages of the disease are minimal, and only highly discriminant features can capture these variations. From the experimental results, as shown in Table 2.16, the features extracted from CNNs provide significantly higher classification accuracy in comparison to the WNDCHRM, and these features are effective and promising for classifying the consecutive stages of knee OA.

Table 2.16 Classification accuracy (%) achieved by the WNDCHRM and pre-trained CNN features

Category	Classification	WNDCHRM	VGG-16 net			VGG-M-128 net			BVLC ref CaffeNet		
			fc7	pool5	conv5-2	fc6	pool5	conv4	fc7	pool5	conv5
Progressive	Grade 0 versus Grade 1	51.5	56.3	61.3	63.5	56.5	63.2	**64.7**	62.0	64.3	63.3
	Grade 0 versus Grade 2	62.6	68.6	74.3	76.7	67.8	75.5	**77.6**	69.6	73.6	73.9
	Grade 0 versus Grade 3	70.6	86.4	91.4	92.4	88.5	90.2	**92.9**	87.9	92.5	91.5
	Grade 0 versus Grade 4	82.8	98.1	98.6	99.3	98.8	99.3	99.2	98.5	**99.4**	99.1
Successive	Grade 1 versus Grade 2	48.8	60.0	64.7	67.3	57.9	63.5	65.3	61.2	**65.8**	62.8
	Grade 2 versus Grade 3	54.5	69.8	76.4	77.0	73.0	77.3	**79.0**	70.3	78.1	77.1
	Grade 3 versus Grade 4	58.6	85.2	88.8	90.0	85.0	90.4	91.2	87.4	**91.6**	91.4
Multi-class	Grade 0 to Grade 2	39.9	51.1	53.4	56.9	51.1	55.0	**57.4**	51.1	54.8	54.4
	Grade 0 to Grade 3	32.0	44.6	48.7	53.9	45.4	50.2	**53.3**	46.9	51.6	50.2
	Grade 0 to Grade 4	28.9	42.6	47.6	53.1	43.8	49.5	**53.4**	44.1	50.8	50.0

Multi-class classifications are performed using linear SVMs with the CNN features (Table 2.16, multi-class). Again, the CNN features outperform WNDCHRM. The classification accuracies obtained using convolutional (conv4, conv5) and pooling (pool5) layers are slightly higher in comparison to fully-connected layer features. There are minimal variations in classification accuracy obtained with the features extracted from VGG-M-128 net and BVLC reference CaffeNet in comparison to VGG-16.

2.5.2.3 Transfer Learning

As a next approach [45], the BVLC CaffeNet [66] and VGG-M-128 [65] networks are fine-tuned using transfer learning to classify knee images. These two smaller networks are chosen because they contain fewer layers and parameters (∼62M), over the much deeper VGG-16, which has ∼138M parameters. The top fully-connected layer of both networks is replaced and the model is retrained on the OAI dataset using backpropagation. The lower-level features in the bottom layers are also updated during fine-tuning. Standard softmax loss is used as the objective for classification, and accuracy layers are added to monitor the training progress. A Euclidean loss layer (mean squared error) is used for the regression experiments.

Experiments and Results. Table 2.17 shows the multi-class classification results for the fine-tuned BVLC CaffeNet. The VGG-16 network is omitted in these experiment since the variation in accuracy among the pre-trained CNNs is small, and fine-tuning VGG-16 is more computationally expensive.

The dataset is split into training (60%), validation (10%) and test (30%) sets for fine-tuning. The right-left flipped knee joint images are included in the training set to increase the number of training samples. The networks are fine-tuned for 20 epochs using a learning rate of 0.001 for the transferred layers, and 0.01 for the newly introduced layers. The performance of fine-tuned BVLC CaffeNet is slightly better than VGG-M-128. Hence, the results of fine-tuning BVLC CaffeNet is only shown here. Figure 2.35 shows the learning curves for training and validation loss, and validation accuracy. The decrease in loss and increase in accuracy shows that the fine-tuning is effective and makes the CNN features more discriminative, which improves classification accuracy (Table 2.16). The features extracted from the fully connected (fc7) layer provide slightly better classification in comparison to pooling (pool5) and convolution (conv5) layers.

Regression using Fine-tuned CNNs. Existing work on automatic assessment of knee OA severity treats it as an image classification problem, assigning each KL grade to a distinct category [1–3, 5]. To date, evaluation of automatic KL grading algorithms has been based on binary and multi-class classification accuracy with respect to these discrete KL grades [3, 5, 6]. Nevertheless, KL grades are not categorical, but rather represent an ordinal scale of increasing severity. Treating them as categorical during evaluation means that the penalty for incorrectly predicting that a subject with grade 0 OA has grade 4 is the same as the penalty for predicting that the same subject has

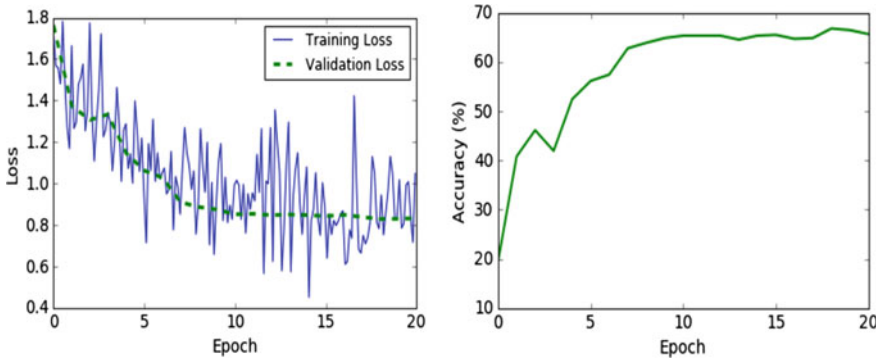


Fig. 2.35 Learning curves: training and validation losses (left), and validation accuracy (right) during fine-tuning

Table 2.17 Classification accuracy (%) achieved with the features extracted from fine-tuned BVLC Net

Classification	Before fine-tuning			After fine-tuning		
	fc7	pool5	conv5	fc7	pool5	conv5
Grade 0 versus Grade 1	62.0	64.3	63.3	63.3	**64.3**	61.9
Grade 0 versus Grade 2	69.6	73.6	73.9	76.3	**77.2**	74.1
Grade 0 versus Grade 3	87.9	92.5	91.5	**96.7**	96.0	96.3
Grade 0 versus Grade 4	98.5	99.4	99.1	**99.8**	99.7	99.7
Grade 1 versus Grade 2	61.2	65.8	62.8	63.3	**66.7**	62.7
Grade 2 versus Grade 3	70.3	78.1	77.1	**85.8**	83.9	83.3
Grade 3 versus Grade 4	87.4	91.6	91.4	**94.4**	93.6	92.6
Grade 0 to Grade 2	51.1	54.8	54.4	**57.4**	57.0	52.0
Grade 0 to Grade 3	46.9	51.6	50.2	**57.2**	56.5	51.8
Grade 0 to Grade 4	44.1	50.8	50.0	**57.6**	56.2	51.8

grade 1 OA. Clearly the former represents a more serious error, yet this is not captured by evaluation measures that treat grades as categorical variables [45]. In this set up, permuting the ordering of the grades has no effect on classification performance. Moreover, the quantisation of the KL grades to discrete integer levels is essentially an artefact of convenience; the true progression of the disease in nature is continuous, not discrete.

The author proposes that it is more appropriate to measure the performance of an automatic knee OA severity assessment system using a continuous evaluation metric like mean squared error. Such a metric appropriately penalises errors in proportion to their distance from the ground truth, rather than treating all errors equally. Directly optimising mean squared error on a training set also naturally leads to the formulation of knee OA assessment as a standard regression problem. Treating it as such provides

Table 2.18 MSE for classification and regression

Classes	WNDCHRM	CNN-Clsf	CNN-Reg	CNN-Reg*
Grade 0–4	2.459	0.836	**0.504**	0.576

the model with more information on the structure and relationship between training examples with successive KL grades. It is demonstrated that the use of regression reduces both the mean squared error and improves the multi-class classification accuracy of the model [45].

The pre-trained BVLC CaffeNet model is fine-tuned using both classification loss (cross entropy on softmax outputs) and regression loss (mean squared error) to compare their performance in assessing knee OA severity. In both cases, the fully connected layer fc7 is replaced with a randomly initialised layer and fine-tuned for 20 epochs, selecting the model with the highest validation performance. The classification network uses a 5D fully connected layer and softmax following the fc7 layer, and the regression network uses a 1D fully connected node with a linear activation.

The models are compared using both mean squared error (MSE) and standard multi-class classification metrics. The mean squared error is calculated using the standard formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where n is the number of test samples, y_i is the true (integer) label and $\hat{y}_i$ is the predicted label. For the classification network the predicted labels y_i are integers and for the regression network they are real numbers. A configuration is tested, where the real outputs are rounded from the regression network to produce integer labels. Table 2.18 shows the MSE for classification using the WNDCHRM and the CNN trained with classification loss (CNN-Clsf), regression loss (CNN-Reg), and regression loss with rounding (CNN-Reg*). Regression loss clearly achieves significantly lower mean squared error than both the CNN classification network and the WNDCHRM features.

To demonstrate that the regression loss also produces better classification accuracy, the classification accuracy from the network trained with classification loss and the network trained with regression loss and rounded labels are compared. Rounding in this case is necessary to allow the use of standard classification metrics. Table 2.19 compares the resulting precision, recall, and F_1 scores. The multi-class (grade 0–4) classification accuracy of the network fine-tuned with regression loss is 59.6%. The network trained using regression loss clearly gives superior classification performance. The author suspects this is due to the fact that using regression loss gives the network more information about the ordinal relationship between the KL grades, allowing it to converge on parameters that better generalise to unseen data.

Table 2.19 Comparison of classification performance using classification (left) and regression (right) losses

Classification	Classification loss			Regression loss		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.53	0.64	0.58	0.57	0.92	0.71
1	0.25	0.19	0.22	0.32	0.14	0.20
2	0.44	0.32	0.37	0.71	0.46	0.56
3	0.37	0.47	0.41	0.78	0.73	0.76
4	0.56	0.54	0.55	0.89	0.73	0.80
Mean	0.43	0.44	0.43	0.61	0.62	0.59

Discussion. The initial approach to quantify knee OA severity used features extracted from pre-trained CNNs. Three pre-trained networks are investigated and it is found that the BVLC reference CaffeNet and VGG-M-128 networks perform best. A linear SVM trained on features from these networks achieved significantly higher classification accuracy (53.4%) in comparison to the previous state-of-the-art (28.9%). The features from pooling and convolutional layers were found to be more accurate than the fully connected layers. Fine-tuning the networks by replacing the top fully connected layer gave further improvements in multi-class classification accuracy.

Previous studies have assessed their algorithms using binary and multi-class classification metrics. The author proposes that it is more suitable to treat KL grades as a continuous variable and assess accuracy using mean squared error. This approach allows the model to be trained using regression loss so that errors are penalised in proportion to their severity, producing more accurate predictions. This approach also has the nice property that the predictions can fall between grades, which aligns with continuous disease progression.

In summary, this section presented two approaches based on the existing pre-trained CNNs for quantifying knee OA severity: first, the CNNs were used for fixed feature extraction and next, the CNNs were fine-tuned using transfer learning. Both the approaches outperformed the previous state-of-the-art, the WNDCHRM classifier, giving promising results. As a next logical step, CNNs are trained from scratch to investigate if this leads to further improvement in quantifying knee OA severity.

2.5.2.4 Training CNNs from Scratch

Training a CNN from scratch (or full training) is challenging and complicated, because it requires a large amount of annotated training data. The learning curves during training should ensure proper convergence to generalise well avoiding overfitting [15]. An alternative to full training is transfer learning, fine-tuning CNNs pre-trained in other domain (for instance ImageNet dataset with natural images) to a target domain, for instance medical domain. However, the knowledge transfer may

be limited by the substantial differences between the source and the target domains, which may mitigate the performance of the fine tuned CNNs. Nevertheless, with sufficient labelled training data and carefully selected hyper-parameters, fully trained CNNs can outperform fine-tuned CNNs and hand-crafted alternatives [12, 15].

Fully trained CNNs have been found to be highly successful in many medical applications [12, 15]. Some of the applications that use fully trained CNNs for musculo-skeletal (including knee) image analysis are knee cartilage segmentation using multi-stream CNNs [16], total knee arthoplasty kinematics by real-time 2D/3D registration using CNN regressors [72], automated skeletal bone age assessment in X-ray images using deep learning [73], and posterior-element fractures detection on spine CT using deep convolutional networks [74]. Motivated by these approaches, CNNs are trained from scratch to quantify knee OA severity using both classification and regression.

Dataset and Preprocessing. The data used for the initial experiments are taken from the baseline OAI dataset. There are 4,446 X-ray images with the KL grade annotations in this dataset. The MOST dataset is included for later experiments and this dataset consists of 2,920 X-ray images with KL grade annotations. Two set of knee joint images are used separately for the experiments: (1) extracted after automatic localisation and (2) extracted after manual annotation of the ROI. This is to compare the quantification performance of the CNNs trained with knee joints from automatic localisation and manual annotation. As a preprocessing step, all the knee joint images are subjected to histogram equalisation for intensity level normalisation. The images were resized to 256×256 pixels for the initial experiments. Later, the input image size is changed to 200×300. This size is chosen to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints. Right-left flip of the knee joint images are used to generate more training data.

Initial Configuration. A CNN is configured with a lightweight architecture with 4 layers of learned weights: 3 convolutional layers and 1 fully connected layer. As the training data set is relatively small, a lightweight architecture is considered with minimal (4.5 million) parameters in comparison to the existing CNNs. Table 2.20 shows the CNN configuration in detail. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. A max pooling layer is included after each convolution stage. The final pooling layer is followed by a fully connected layer (fc4), and a softmax dense layer (fc5) with an output shape 5 for the multi-class classification of (0–4) ordinal KL grades. A drop out layer with a drop out ratio of 0.5 is included after the fully connected layer (fc5) to avoid overfitting. The input images are of size 256×256 pixels and fed to the network after sub-sampling by a factor of 2. So, the input size is 128×128 pixels.

Training Process and Initial Results. The network parameters are trained from scratch with the knee joint images as training samples and the KL grades (0, 1, 2, 3 or 4) as labels. To start, the knee joint images extracted manually from the radiographs of the OAI dataset are used. The dataset is split into training (70%) and test (30%) sets. The validation (10%) data is taken from the training set. The network is trained to

Table 2.20 Initial CNN configuration

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 128 × 128
maxPool1	–	3 × 3	3	32 × 42 × 42
conv2	96	7 × 7	1	96 × 42 × 42
maxPool2	–	3 × 3	3	96 × 14 × 14
conv3	128	3 × 3	1	128 × 14 × 14
maxPool3	–	3 × 3	2	128 × 4 × 4
fc4	–	–	–	2048
fc5	–	–	–	5

minimise categorical cross entropy for multi-class classification. *Stochastic gradient descent* (SGD) is used with default parameters: decay = $1e^{-6}$, momentum = 0.9, and nesterov = True and the initial learning rate is set to 0.0001. The networks are trained with fixed learning rate in the initial experiments. The Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$ is tested, instead of SGD for the later experiments. The benefits of the Adam optimiser are that it uses adaptive learning rates and provides faster convergence.

This network achieves a multi-class classification accuracy of 44.7% on the test data. The mean-squared error is 1.75. Table 2.21 shows the classification results: precision, recall, and F_1 score of the initial configuration. The results show that the classification performance is low and the mean-squared error is high. These are initial results and the hyper-parameters of this network are tuned to improve the classification performance. Further, the number of convolutional layers, convolutional-pooling stages, the number of convolutional kernels, kernel sizes and other parameters are experimented.

The terms 'parameters' and 'hyper-parameters' in machine learning are often used interchangeably, but there is a difference between them. Parameters are learned by a classifier or a machine learning model from the training data, for instance weights or coefficients of the independent variables. Hyper-parameters are the settings used to optimise the performance of a classifier or a model and they are not fit based on the training data. The hyper-parameters for a CNN include the number and size of the hidden layers, learning rate and its decay, drop out regularisation, gradient clipping threshold and other settings.

Tuning Hyper-parameters. After the initial CNN configuration giving low classification accuracy (44.7%), as a first step the depth of the network is increased. A convolutional layer and a pooling layer are included. This increases the number of layers with learned weights to 5 layers: 4 convolutional layers and 1 fully connected layer. SGD with default parameters: decay = $1e^{-6}$, momentum = 0.9, and nesterov = True, is used for training this network. Learning rates from 0.0001 to 0.01 with an incremental increase by a factor 10 are tested, and the learning rate 0.001 is found to be the best. After experimenting with the convolutional kernel size, the number

Table 2.21 Classification results of the initial CNN configuration

Grade	Precision	Recall	F_1 score
0	0.45	0.92	0.60
1	0.24	0.07	0.11
2	0.49	0.18	0.26
3	0.50	0.39	0.44
4	1.00	0.01	0.02
Mean	0.45	0.45	0.37

of kernels in the convolutional layer, the number of outputs of the fully connected layer and other parameters, the final architecture in this configuration is obtained. Table 2.22 shows the CNN architecture in detail.

After 20 epochs of training, this network gave a multi-class classification accuracy of 55.2% with a mean-squared error 0.803 on the validation data. After 35 epochs the network achieves the best results for this configuration with a classification accuracy of 60.4% and mean-squared error 0.838. Table 2.23 shows the classification results: precision, recall, and F_1 score of this network. There is an improvement in the overall classification results in comparison to the previous results (Table 2.21). Figure 2.36 shows the learning curves with increase in the training and validation accuracies, and decrease in the training and validation losses whilst training this network. It can be observed from the learning curves (Fig. 2.36), after 32 epochs there is an increase in validation loss with decrease in training loss and also there is no further increase in validation accuracy whilst training accuracy increases: the network is starting to overfit. A drop out regularisation by a ratio of 0.5 is included after the fully connected layer (fc5) to mitigate overfitting. Also, data augmentation is used to increase the training samples by including the right-left flip of the knee joints and this doubles the number of training samples. Drop out regularisation after convolutional layers and fully connected layers, and l2-norm weight regularisations are used to further mitigate overfitting in the next set of experiments.

Next, the depth of the network is further increased by increasing the number of layers with learned weights, continuing the experimentation with the other associated hyper-parameters. Up to 5 convolutional-pooling stages followed by two fully connected layers are tested. The classification accuracy with 4 convolutional-pooling stages is 60.8% and with 5 convolutional-pooling stages is 61%.

Previous networks use a single convolutional layer followed by a pooling layer. Next, cascaded convolutional layers are used in a convolution-pooling stage like VGG-16 model. Each convolutional layer is followed by a ReLU activation. Figure 2.24 shows the CNN architecture that gives the best results in this approach. This network gives a classification accuracy of 60.1% with a mean-squared error 0.838.

Inspired by the success of VGG networks [53], a network with cascaded convolutional layers of uniform (3×3) kernel size and (2×2) max pooling with stride 2 is trained, and the hyper-parameters are tuned. This network gives a classification

Table 2.22 CNN architecture (CNN-1) after tuning hyper-parameters

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 128 × 128
maxPool1	–	3 × 3	2	32 × 63 × 63
conv2	96	5 × 5	1	64 × 63 × 63
maxPool2	–	3 × 3	2	64 × 31 × 31
conv3	128	3 × 3	1	128 × 31 × 31
maxPool3	–	3 × 3	2	128 × 15 × 15
conv4	256	3 × 3	1	256 × 15 × 15
maxPool4	–	3 × 3	2	256 × 7 × 7
fc5	–	–	–	1024
fc6	–	–	–	5

Table 2.23 Classification results after tuning hyper-parameters

Grade	Precision	Recall	F_1 score
0	0.57	0.90	0.70
1	0.31	0.11	0.16
2	0.64	0.45	0.53
3	0.74	0.77	0.76
4	0.86	0.72	0.78
Mean	0.58	0.60	0.57

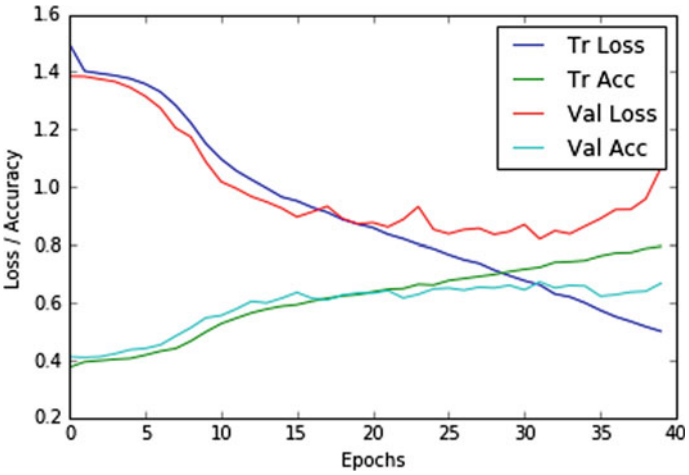


Fig. 2.36 Learning curves: training and validation losses, and accuracies of the fully trained CNN

Table 2.24 CNN architecture (CNN-2) after tuning hyper-parameters

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 128 × 128
maxPool1	–	3 × 3	2	32 × 63 × 63
conv2	64	5 × 5	1	64 × 63 × 63
maxPool2	–	3 × 3	2	64 × 31 × 31
conv3-1	64	3 × 3	1	64 × 31 × 31
conv3-2	64	3 × 3	1	64 × 31 × 31
maxPool3	–	3 × 3	2	64 × 15 × 15
conv4-1	96	3 × 3	1	96 × 15 × 15
conv4-2	96	3 × 3	1	96 × 15 × 15
maxPool4	–	3 × 3	2	96 × 7 × 7
fc5	–	–	–	1024
fc6	–	–	–	5

accuracy of 57.5% with a mean-squared error 0.961. There is no further improvement in the classification results in comparison to the previous results.

Training Off-the-shelf CNNs from Scratch. Earlier, the widely used off-the-shelf CNNs such as BVLC reference CaffeNet [66, 67] (which is very similar to the AlexNet model [14]), VGG-M-128 network [65], and VGG-16 network [53] were fine-tuned for knee images classification. The pre-trained VGG-16 network has ∼138 million free parameters, and the other networks, Alexnet with ∼62 million and the VGG-M-128 with ∼26 million parameters, are relatively simple. Training these networks, in particular VGG-16, from scratch is computationally very expensive due to the depth and the number of free parameters. Previously trained CNNS have relatively fewer parameters (∼4 to 6 million) to suit the relatively small dataset with a few thousand of training examples.

Next, CNNs are fully trained using the AlexNet and the VGG-M-128 architectures. This is to compare the classification performance of these networks to the previously trained networks from scratch. Table 2.25 shows the AlexNet architecture in detail. The convolutional layers conv1 and conv2 in this network are followed by Relu and batch normalisation layers. The two fully connected layers (fc6) and (fc7) are followed by a drop out regularisation by a ratio 0.5. This network was pre-trained for 1,000 classes in the ImageNet [75] dataset. The output of the last fully connected layer (fc8) is replaced with a 5 output dense layer for multi-class knee OA image classification. This network is trained using SGD with default parameters. Learning rates from 0.00001 to 0.01 with an incremental increase by a factor 10 are tested. The learning rate set at 0.001 gives the best results.

The fully trained AlexNet gives a classification accuracy of 57.2% with a mean-squared error 0.741. Table 2.26 shows the classification results; precision, recall, and F_1 score of the fully trained AlexNet model. These results show that the classification

Table 2.25 AlexNet architecture

Layer	Kernels	Kernel size	Strides	Output shape
conv1	96	11 × 11	4	96 × 64 × 64
maxPool1	–	3 × 3	2	96 × 31 × 31
conv2	256	5 × 5	1	256 × 31 × 31
maxPool2	–	3 × 3	2	256 × 15×15
conv3	384	3 × 3	1	384 × 15×15
conv4	384	3 × 3	1	384 × 15×15
conv5	256	3 × 3	1	256 × 15×15
maxPool5	–	3 × 3	2	256 × 7×7
fc6	–	–	–	4096
fc7	–	–	–	4096
fc8	–	–	–	5

Table 2.26 Classification results of the fully trained AlexNet

Grade	Precision	Recall	F_1 score
0	0.65	0.61	0.63
1	0.29	0.36	0.32
2	0.59	0.55	0.57
3	0.75	0.73	0.74
4	0.77	0.79	0.78
Mean	0.59	0.57	0.58

accuracy achieved by the fully trained AlexNet is low (57.2%) in comparison to the accuracy (60.8%) achieved by previous networks. Moreover, this network is overfitting. This is evident from the learning curves (Fig. 2.37) obtained whilst training this network. After 30 epochs, the learning curves show an increase in validation loss whilst the training loss is decreasing and there is no improvement in the validation accuracy whilst the training accuracy keeps increasing. The reason for overfitting is the number of training samples in the dataset (∼10,000) is very low in comparison to the number of free parameters (∼62 million) in AlexNet. This model was originally developed and trained on datasets like ImageNet [75] that consists of more than ∼1.2 million images. There are two fully connected layers with 4,096 outputs in the AlexNet and these layers contribute to more than 95% of the total free parameters in this network. Next, a relatively simple architecture (VGG-M-128) is investigated for the knee OA images classification.

The VGG-M-128 network is a simplified model of the AlexNet [53]. The last fully connected layer (fc7) of AlexNet has 4,096 outputs. The number of fc7 outputs is reduced to 128 in VGG-M-128. This reduces the number of free parameters and this network contains (∼26 million) parameters in total. The AlexNet configuration is retained in the VGG-M-128 network with a few changes in the architecture. The

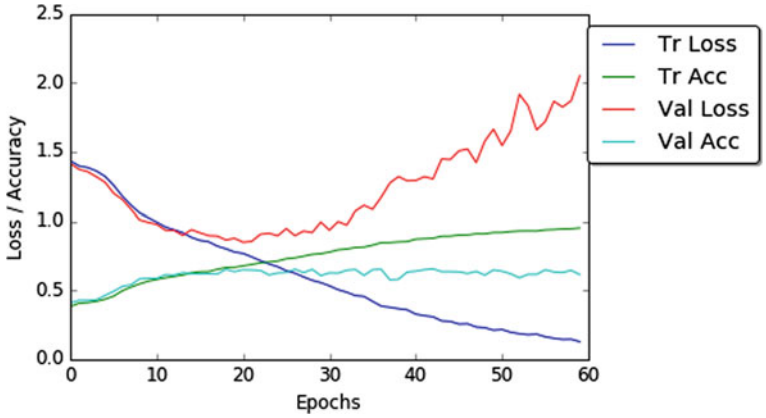


Fig. 2.37 Learning curves: training and validation losses, and accuracies of the fully trained AlexNet

Table 2.27 VGG-M-128 architecture

Layer	Kernels	Kernel size	Strides	Output shape
conv1	96	7 × 7	2	96 × 128 × 128
maxPool1	–	3 × 3	2	96 × 63 × 63
conv2	256	5 × 5	1	256 × 32 × 32
maxPool2	–	3 × 3	2	256 × 15 × 15
conv3	512	3 × 3	1	512 × 15 × 15
conv4	512	3 × 3	1	512 × 15 × 15
conv5	512	3 × 3	2	512 × 8 × 8
maxPool5	–	3 × 3	2	512 × 3 × 3
fc6	–	–	–	4096
fc7	–	–	–	128
fc8	–	–	–	5

kernel size of the first convolutional layer is reduced to (7 × 7) and the stride is reduced to 2. The number of filters is fixed to 512 in the conv3, conv4, and conv5 layers. Table 2.27 shows the architecture details. This network parameters are trained from scratch using SGD with default parameters: decay = $1e^{-6}$, momentum = 0.9, and nesterov = True. The learning rate is fixed to 0.001 after testing different rates like before.

This network gives a classification accuracy of 56.3% and the mean-squared error is 0.685. Table 2.28 shows the classification results of this network. The results show a slightly lower classification accuracy (56.3%) in comparison to the previous results. There is no significant difference in the precision, recall, and F_1 score of this network in comparison to the AlexNet classification results (Table 2.26). This network is also overfitting like the AlexNet. This is evident from the learning curves (Fig. 2.38) of

Table 2.28 Classification results of the fully trained VGG-M-128

Grade	Precision	Recall	F_1 score
0	0.66	0.65	0.66
1	0.27	0.42	0.33
2	0.62	0.46	0.53
3	0.77	0.69	0.72
4	0.87	0.73	0.79
Mean	0.60	0.56	0.58

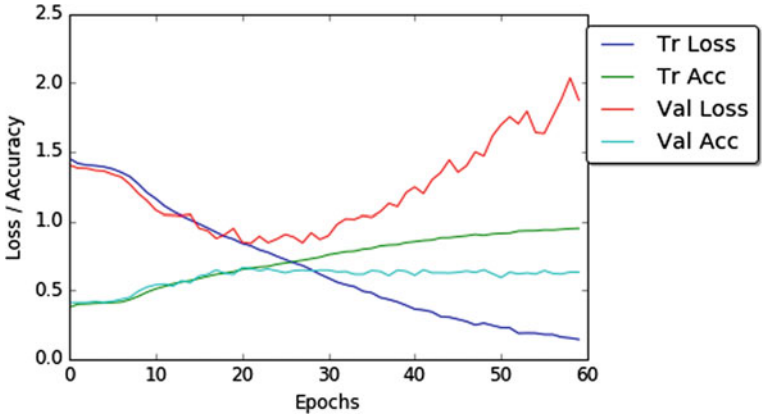


Fig. 2.38 Learning curves: training and validation losses, and accuracies of the fully trained VGG-M-128 network

this network. The learning curves show increase in the validation loss after 30 epochs and the validation accuracy remains almost the same. The drop out regularisations after the fully connected layers fc6 and fc7 are not able to fully mitigate overfitting. The reason for overfitting remains the same as for AlexNet. The number of training samples is very low even for the number of free parameters in this network ($\sim$26 million).

Best Performing CNN for Classification. After experimenting with different configurations, the network in Table 2.29 is found to be the best for classifying knee images. This network is similar to the previous configuration (Table 2.24), but with slight variations. The network contains five layers of learned weights: four convolutional layers and a fully connected layer. The total number of free parameters in the network is $\sim$5.4 million. Each convolutional layer in the network is followed by batch normalisation and a ReLU activation layer. After each convolutional stage there is a max pooling layer. The final pooling layer (maxPool4) is followed by a fully connected layer (fc5) and a softmax dense (fc6) layer. To avoid overfitting, a drop out layer with a drop out ratio of 0.25 is included after the last convolutional (conv4) layer and a drop out layer with a drop out ratio of 0.5 after the fully connected

Table 2.29 Best performing CNN for classifying the knee images

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 100 × 150
maxPool1	–	3 × 3	2	32 × 49 × 74
conv2	64	5 × 5	1	64 × 49 × 74
maxPool2	–	3 × 3	2	64 × 24 × 36
conv3	96	3 × 3	1	96 × 24 × 36
maxPool3	–	3 × 3	2	96 × 11 × 17
conv4	128	3 × 3	1	128 × 11 × 17
maxPool4	–	3 × 3	2	128 × 5 × 8
fc5	–	–	–	1024
fc6	–	–	–	5

layer (fc5). Also, a L2-norm weight regularisation penalty of 0.01 is applied in the last two convolutional layers (conv3 and conv4) and the fully connected layer (fc5). Applying a regularisation penalty to other layers increases the training time whilst not introducing significant variation in the learning curves. The network is trained to minimise categorical cross-entropy loss using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. The inputs to the network are knee images of size 200 × 300. This size is selected to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints.

First, this network is trained using the OAI dataset like the previous network trainings. This network achieves a classification accuracy of 61% with a mean-squared error 0.861. Next, training samples are included from the MOST dataset. This network achieves a classification accuracy of 61.8% with a mean-squared error 0.735 for the combined OAI-MOST dataset. There is a slight increase in the classification accuracy (0.8%) and decrease in the mean-squared error (0.126). Table 2.30 shows the classification results: precision, recall, and F_1 score of this network for the combined OAI-MOST dataset. Figure 2.39 shows the learning curves whilst training this network. The learning curves show proper convergence of the training and validation losses with consistent increase in the training and validation accuracies till they reach constant values.

To sum up, a high classification accuracy (61%) is achieved with the CNN (Table 2.29) trained from scratch and outperform the VGG-M-128 and the AlexNet trained from scratch. The fully trained AlexNet gives a classification accuracy of 57.2% and VGG-M-128 gives an accuracy of 56.3%. The classification results of the methods proposed in this section and the previous state-of-the-art are compared in the next section.

Classification Results. The classification results of the fully trained network is compared to WNDCHARM, the multi purpose medical image classifier [4, 5, 60] that gave the previous best results for automatically classifying knee OA X-ray images,

Table 2.30 Classification results of the best performing fully trained CNN

Grade	Precision	Recall	F_1 score
0	0.65	0.83	0.73
1	0.30	0.10	0.14
2	0.51	0.60	0.55
3	0.77	0.69	0.73
4	0.87	0.70	0.78
Mean	0.59	0.62	0.59

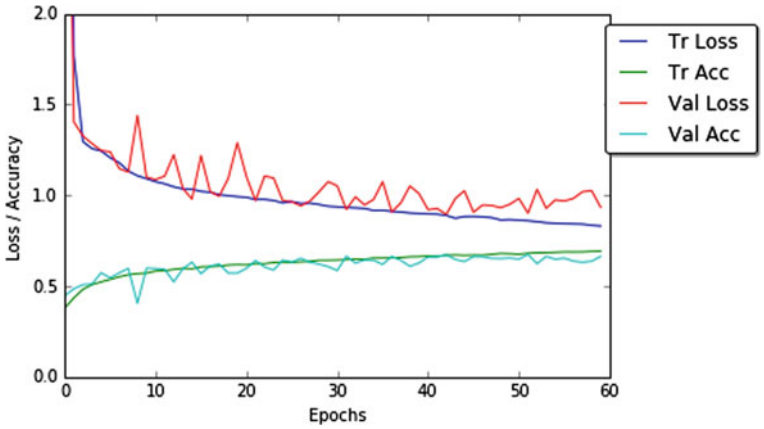


Fig. 2.39 Learning curves: training and validation losses, and accuracies of the best performing fully trained CNN

Table 2.31 Classification results of the proposed methods and the existing methods

Method	Test data	Accuracy (%)	Mean-squared error
WNDCHRM	OAI and MOST	34.8	2.112
Fine-tuned BVLC CaffeNet	OAI	57.6	0.836
Fully trained CNN	OAI	61	0.861
Fully trained CNN	OAI and MOST	**61.8**	**0.735**

and to previous results (Table 2.17) on fine-tuning BVLC reference CaffeNet for this task (Sect. 2.5.2.3). WNDCHARM is trained with the data taken from the OAI and MOST datasets.

Table 2.31 shows the multi-class classification accuracy and mean-squared error of the fine-tuned BVLC CaffeNet, the network trained from scratch and WNDCHARM for the OAI and MOST datasets. The results show that the network trained from scratch for classifying knee OA images clearly outperforms WNDCHARM.

This shows learning feature representations using CNNs for fine-grained knee OA images classification is highly effective and a better approach in comparison to using a combination of hand-crafted features in WNDCHARM. The other reason for low classification accuracy of WNDCHARM is that it uses only a balanced dataset for training. Both the OAI and MOST datasets are very unbalanced and in particular the number of knee images available in KL grade 4 is very small, ∼5% in total.

Moreover, these results show an improvement over previous methods that used fine-tuned off-the-shelf networks such as VGG-M-128 and the BVLC Reference CaffeNet for classifying knee OA X-ray images through transfer learning. These improvements are due to the lightweight architecture of the network trained from scratch with less (∼5.4 million) free parameters in comparison to 62 million free parameters of BVLC CaffeNet for the small amount of training data available. The off-the-shelf networks are trained using a large dataset like ImageNet containing millions of images, whereas the dataset used in this experiment contains much fewer (∼10,000) training samples. Furthermore, the results show an increase in classification accuracy from 61 to 61.8% when the MOST dataset is included in the training set. This result is promising and it shows that with more training data the CNN performance maybe further improved. Next, the use of regression by fully trained CNNs is investigated to improve the quantification performance.

Training CNNs for Regression. CNNs are trained from scratch to classify knee images in the previous approach. The outcomes are ordinal KL grades (0, 1, 2, 3 or 4) that quantify knee OA severity. CNNs are trained for regression in the next approach. This is to assess knee OA severity in a continuous scale (0–4). The author argued earlier (Sect. 2.5.2.2) that it is more appropriate to assess knee OA in a continuous scale as knee OA is progressive in nature, not discrete [45]. The existing CNNs are fine-tuned to quantify knee OA severity using regression.

Initial Configuration. A CNN is trained for regression using almost the same architecture (Table 2.29) that gave the highest multi-class classification accuracy previously. The last fully connected layer (fc6) with softmax activation and an output shape 5 for multi-class classification is replaced with a linear activation with an output shape of 1 for regression. The CNN is trained to minimise mean-squared error using the Adam optimiser with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. Like before, the inputs to the network are images of size 200 × 300. The data for training is taken from both the OAI and the MOST datasets. Both these datasets contain discrete KL grade (0, 1, 2, 3 or 4) annotations for the knee joints. These labels are used in the previous approach to train classifiers. However, there is no ground truth for KL grades on a continuous scale for either of these datasets to train a network directly for regression output. Hence, the discrete KL grades are used as labels to train CNNs for regression.

Initial Results. This CNN gives a mean-squared error of 0.654 on the test data after training. In comparison to the mean-squared error achieved by the classifier (0.898) with almost the same architecture, there is definitely an improvement in the quantification using regression. The performance metrics: accuracy, precision, recall, and F_1

score are computed for the regression results by rounding the predicted continuous grade to the next integer value. Rounding, in this case, is necessary to allow the use of standard classification metrics and to compare the performances of classification and regression. Table 2.32 shows the precision, recall, and F_1 score for regression. In comparing these results to the previous classification results (Table 2.30), there is a decrease in precision, recall, and F_1 score. The classification accuracy achieved by regression is 36.9% with a mean-squared error 0.75. From these results it is evident that the regression performance is low in this initial configuration. Next, the hyper parameters of this network are tuned to improve the regression performance.

Tuning the Hyper-parameters. The experiment is continued by varying the number of layers with learned weights in the architecture, the number of convolutional-pooling stages, the number of kernels and kernel sizes in the convolutional layers and the regularisations to avoid overfitting. The architecture in Table 2.33 is found to be the best for quantifying knee OA severity using regression. This network contains seven layers of learned weights: six convolutional layers and a fully connected layer.

Table 2.32 Results of the initial network trained for regression after rounding the predicted continuous grades

Grade	Precision	Recall	F_1 score
0	0.78	0.18	0.29
1	0.24	0.83	0.37
2	0.49	0.32	0.39
3	0.63	0.42	0.50
4	0.57	0.20	0.30
Mean	0.57	0.37	0.36

Table 2.33 Best performing CNN for regression of the knee images

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11×11	2	$32 \times 100 \times 158$
maxPool1	–	3×3	2	$32 \times 49 \times 74$
conv2	64	5×5	1	$64 \times 49 \times 74$
maxPool2	–	3×3	2	$64 \times 24 \times 36$
conv3-1	64	3×3	1	$64 \times 24 \times 36$
conv3-2	64	3×3	1	$64 \times 24 \times 36$
maxPool3	–	3×3	2	$64 \times 11 \times 17$
conv4-1	128	3×3	1	$96 \times 11 \times 17$
conv4-2	128	3×3	1	$96 \times 11 \times 17$
maxPool4	–	3×3	2	$96 \times 5 \times 8$
fc5	–	–	–	1024
fc6	–	–	–	1

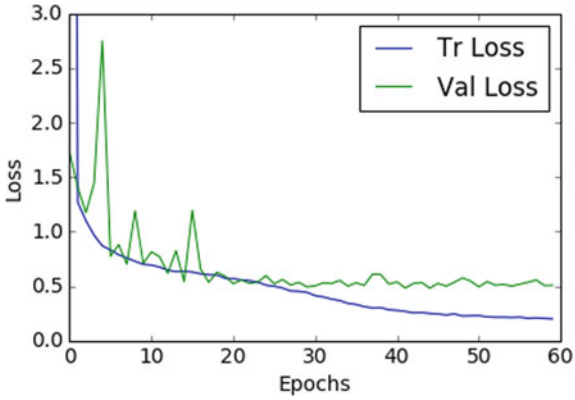


Fig. 2.40 Learning curves: training and validation losses for the best performing CNN for regression

This network has ∼5.6 million free parameters in total. Each convolutional layer is followed by batch normalisation and a ReLU layer. The last pooling layer (maxPool4) is followed by two dense layers: fc5 with ReLU and fc6 with linear activations. A drop out layer with a drop out ratio 0.5 is added after fc6. The network is trained to minimise the mean-squared error using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. The network is trained with knee images taken from the OAI and the MOST datasets. Figure 2.40 shows the learning curves: training and validation losses whilst training this network. The learning curves show convergence in the losses.

Comparison of Classification and Regression Results. The best performing CNN for regression gives a mean-squared error of 0.574. After rounding the continuous grade predictions, this network achieves a multi-class classification accuracy of 54.7% and the mean-squared error is 0.661. Table 2.34 shows the accuracy and mean-squared error for the fully trained CNN for classification and regression. The results show that the multi-class classification accuracy calculated after rounding the output is low for CNN-regression. The main reason for this likely is training the regression network with ordinal labels instead of continuous labels. There is also a decrease in accuracy due to the rounding of regression output and the rounding is necessary to compute standard classification metrics. On the other hand, the mean-squared error of the fully trained CNN for regression is low in both the cases before rounding (0.574) and after rounding (0.661) in comparison to the fully trained CNN

Table 2.34 Comparison of classification and regression results

Method	Accuracy (%)	MSE (before rounding)	MSE (after rounding)
CNN-classification	**61.8**	0.735	–
CNN-regression	54.7	**0.574**	0.661

Table 2.35 Comparison of the regression and classification performances

Grade	Regression			Classification		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.70	0.71	0.70	0.65	0.83	0.73
1	0.29	0.42	0.34	0.30	0.10	0.14
2	0.52	0.39	0.45	0.51	0.60	0.55
3	0.67	0.51	0.58	0.77	0.69	0.73
4	0.58	0.55	0.57	0.87	0.70	0.78
Mean	0.57	0.55	0.55	0.59	0.62	0.59

for regression. Table 2.35 shows the precision, recall, and F_1 score of the rounded regression output and the classification output. These results show that the network trained with classification loss outperforms the regression loss. The reason for this is again likely the lack of continuous KL grade ground truth to train a CNN directly for regression output.

To sum up, training a CNN from scratch for regression output gives low mean-squared error. The lack of ground truth affects the performance of the regression. To overcome this drawback, in the next approach multi-objective convolutional learning is investigated to quantify knee OA severity.

2.5.2.5 Multi-objective Convolutional Learning

In general, assessing knee OA severity is based on the multi-class classification of knee images and assigning KL grade to each distinct category [3–6]. The author previously argued that assigning a continuous grade (0–4) to knee images through regression is a better approach for quantifying knee OA severity as the disease is progressive in nature. However, there is no ground truth i.e. KL grades on a continuous scale to train a network directly for regression output. Therefore, the networks are trained using multi-objective convolutional learning [76–80] to optimise a weighted-ratio of two loss functions: categorical cross-entropy and mean-squared error. Mean squared error gives the network information about the ordering of grades, and cross entropy gives information about the quantisation of grades. Intuitively, optimising a network with two loss functions provides a stronger error signal and it is a step to improve the overall quantification, considering both classification and regression results.

Initial Configuration. The same architecture of the best performing CNN is used for classification (Table 2.29) as an initial configuration to jointly train a CNN for classification and regression outputs. Table 2.36 and Fig. 2.41 shows the configuration details of the initial configuration. The network has five layers with learned weights: four convolutional layers and a fully connected layer. The total free parameters in the network are ∼5.4 million. The last fully connected layer (fc5) is followed by

two dense layers with softmax and linear activations for simultaneous multi-class classification and regression outputs. Drop out layers with a drop out ratio 0.25 are included after the conv4 layer and a drop out ratio 0.5 after the fc6 layer to avoid overfitting. In addition to this, a L2-norm weight regularisation penalty of 0.01 is applied in conv3, conv4 and fc6 layers to avoid overfitting. Applying a regularisation penalty to other layers did not introduce significant variations in the learning curves. Unlike the previous approaches, this network is trained to minimise a weighted ratio of two loss functions: categorical cross-entropy and mean-squared error. After testing different values from 0.2 to 0.6 for the weight of regression loss, a ratio of 0.5 is fixed, as this ratio gives optimal results.

The input to the network are knee images of size 200 × 300. The knee images taken from the combined OAI-MOST dataset is used for training this network. The same train (70%) and test (30%) split are maintained from the previous experiments

Table 2.36 Initial configuration to jointly train a CNN for classification and regression outputs

Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 100 × 150
maxPool1	–	3 × 3	2	32 × 49 × 74
conv2	64	5 × 5	1	64 × 49 × 74
maxPool2	–	3 × 3	2	64 × 24 × 36
conv3	96	3 × 3	1	128 × 24 × 36
maxPool3	–	3 × 3	2	128 × 11 × 17
conv4	128	3 × 3	1	256 × 11 × 17
maxPool4	–	3 × 3	2	256 × 5 × 8
fc5	–	–	–	1024
fc6-Clsf	–	–	–	5
fc6-Reg	–	–	–	1

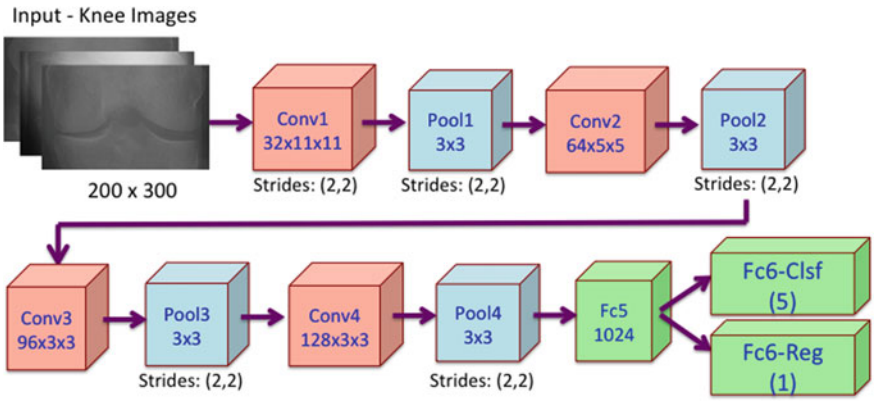
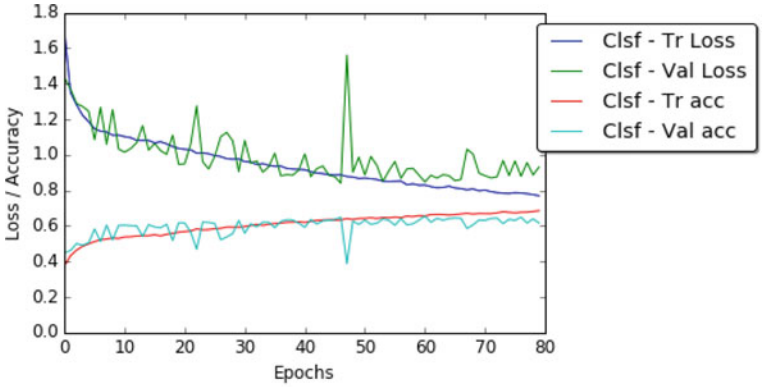


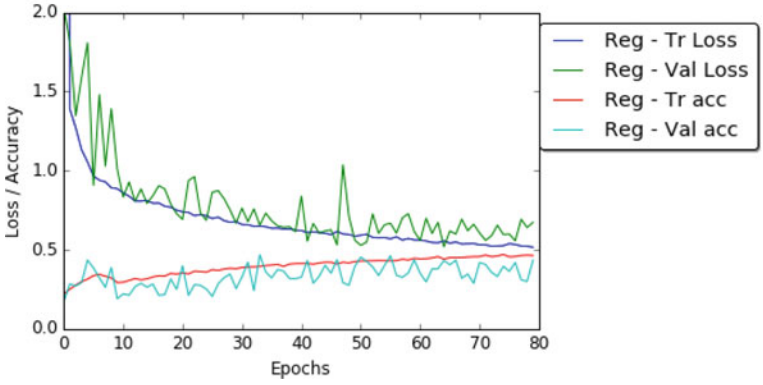
Fig. 2.41 Initial configuration to jointly train a CNN for classification and regression outputs

to make valid comparisons of the quantification results from the different methods. The right-left flip of the knee images is included to increase the training data. A validation split of 20% from the training data is used. This network is trained using the Adam optimise with default parameters: initial learning rate (α) = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, as it gives faster convergence in comparison to the standard SGD.

Initial Results. Figure 2.42 shows the learning curves obtained whilst jointly training the CNN for classification and regression outputs. The learning curves show convergence in the validation and training losses with improvement in validation and classification accuracies. The jointly trained CNN with the initial configuration gives a classification accuracy of 60.8% with mean-squared error 0.795 for the classification outputs and 0.652 for the regression outputs. These results do not show improvement from the previous results. Previously, the network with the same con-



(a) Classification



(b) Regression

Fig. 2.42 Learning curves for **a** classification and **b** regression in jointly trained CNN

figuration gave a classification accuracy of 61.8% and a mean-squared error 0.735 (Table 2.30) when trained to minimise only the classification loss. The same configuration after training to minimise only with the regression loss gave a mean squared error of 0.654 (Table 2.32). This configuration is optimal to minimise classification loss as it gave the highest classification accuracy (61.8%). However, this configuration is not optimal for regression as it gives a high mean-squared error (0.654). Next, the number of layers with learned weights and other hyper-parameters in this configuration are varied to find a good architecture that will give improved results for both classification and regression outputs.

Tuning Hyper-parameters. Cascaded convolutional stages are included in the next configuration in an attempt to improve the regression outputs. The CNNs with cascaded convolutional stages gave best results for regression (Table 2.33) in the previous approach. Table 2.37 shows the network details. This network contains six layers of learned weights: five convolutional layers and a fully connected layer. The total free parameters in this network are ∼7.8 million. The other settings remain the same from the previous network and the same training procedure is followed.

This network gives a multi-class classification accuracy of 62.9% and the mean-squared error is 0.754 for the classification output and 0.583 for the regression output. These results show improvement in the quantification performance in comparison to the previous results. Tuning the hyper parameters improves both the classification and the regression outcomes. Next, the depth of the architecture is increased and other related hyper parameters are tuned to investigate further improvement in the classification and regression outputs.

Best Performing Jointly Trained CNN. The best configuration (Table 2.38) is obtained after experimenting with different settings for jointly training a CNN for classification and regression outputs. This network has eight layers with learned weights: seven convolutional layers and a fully connected layer. This network has

Table 2.37 Jointly trained network for classification and regression outputs

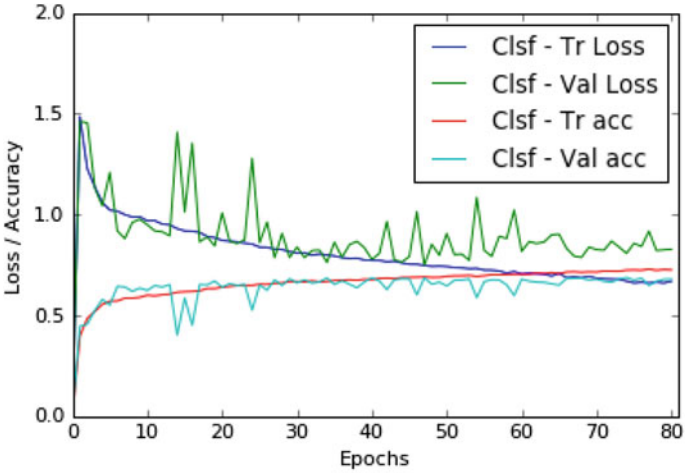
Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 100 × 150
maxPool1	–	3 × 3	3	32 × 33 × 50
conv2-1	64	3 × 3	1	64 × 33 × 50
conv2-2	64	3 × 3	1	64 × 33 × 50
maxPool2	–	3 × 3	2	64 × 16 × 24
conv3-1	96	3 × 3	1	96 × 16 × 24
conv3-2	96	3 × 3	1	96 × 16 × 24
maxPool3	–	3 × 3	2	96 × 7 × 11
fc4	–	–	–	1024
fc5-Clsf	–	–	–	5
fc5-Reg	–	–	–	1

Table 2.38 Jointly trained network for classification and regression outputs

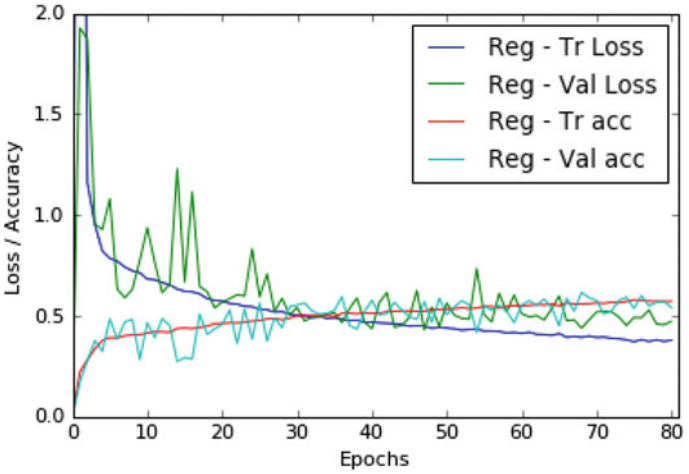
Layer	Kernels	Kernel size	Strides	Output shape
conv1	32	11 × 11	2	32 × 100 × 150
maxPool1	–	3 × 3	2	32 × 49 × 74
conv2-1	64	3 × 3	1	64 × 49 × 74
conv2-2	64	3 × 3	1	64 × 49 × 74
maxPool2	–	3 × 3	2	64 × 24 × 36
conv3-1	96	3 × 3	1	96 × 24 × 36
conv3-2	96	3 × 3	1	96 × 24 × 36
maxPool3	–	3 × 3	2	96 × 11 × 17
conv4-1	128	3 × 3	1	128 × 11 × 17
conv4-2	128	3 × 3	1	128 × 11 × 17
maxPool4	–	3 × 3	2	128 × 5 × 8
fc5	–	–	–	512
fc6-Clsf	–	–	–	5
fc6-Reg	–	–	–	1

∼2.9 million free parameters in total. This is a lightweight architecture with minimal parameters in comparison to the previous networks and the existing off-the-shelf CNNs. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. The fc5 layer is followed by two dense layers with softmax and linear activations for multi-class classification and regression outputs. To avoid overfitting, drop out with ratio 0.3 is included after the last fully connected (fc5) layer. Also, a L2 weight regularisation penalty of 0.01 is applied to all the convolutional and fully connected layers except the first two convolutional layers. This network is trained to minimise a weighted ratio of two loss functions: categorical cross-entropy and mean-squared error. This network is trained using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}^{-8}$.

Jointly Trained CNN Results. Figure 2.43 shows the learning curves obtained whilst jointly training the CNN for classification and regression outputs. The learning curves show convergence to the minimum of the validation and training losses with improvement in validation and classification accuracies. The combined OAI-MOST dataset is used to compute these results. The same train-test split is maintained from the previous experiments. This jointly trained network gives a multi-class classification accuracy of 64.6% with a mean-squared error 0.685 for the classification outputs and 0.507 for the regression outputs. Table 2.39 shows the precision, recall, and F_1 score of this network. There is an improvement in the results: the classification accuracy increases to 64.6% from the initial configuration (60.8%), the mean-squared error for regression decreases to 0.507 from the initial configuration (0.652). Increasing the depth of the architecture by including more layers with learned weights to the initial configuration and tuning the other hyper-parameters improves both the classification



(a) Classification



(b) Regression

Fig. 2.43 Learning curves for **a** classification and **b** regression in jointly trained CNN

and regression results. Intuitively, providing a stronger error signal using both the classification and regression loss to the network allow to fit more parameters.

Results Comparison. The results of the jointly trained CNN are compared to the previous CNNs trained separately for classification and regression outputs. Table 2.40 shows the multi-class classification accuracy and mean-squared error of the jointly trained CNN and the separately trained CNNs for classification and regression outputs. There is an improvement in the classification accuracy and also the mean-squared error decreases for the joint training. These results show that the network

Table 2.39 Results of the best performing jointly trained CNN for classification and regression outputs

Grade	Precision	Recall	F_1 score
0	0.68	0.85	0.75
1	0.34	0.07	0.12
2	0.53	0.63	0.57
3	0.74	0.77	0.75
4	0.86	0.81	0.84
Mean	0.62	0.65	0.60

Table 2.40 Comparison of results from jointly trained CNN and individually trained CNNs for classification and regression results

Method	Clsf-accuracy (%)	Clsf-MSE	Reg-MSE
CNN-classification	61.8	0.735	–
CNN-regression	54.7	–	0.574
Jointly trained CNN	**64.6**	**0.685**	**0.507**

jointly trained for classification and regression learns a better representation in comparison to the previous network trained separately for classification and regression outputs.

In summary, CNNs are trained from scratch to quantify knee OA severity using three approaches: classification, regression and jointly training for simultaneous classification and regression. From the results it is evident that the joint training outperforms both the individual training for classification and regression outputs. This supports the hypothesis that training a CNN for optimising a weighted ratio of two loss functions can improve the overall quantification of knee OA severity.

Error analysis. A confusion matrix and the area under curve (AUC) after plotting the receiver operating characteristics are computed to perform an error analysis on the classification of the knee images by the jointly trained CNN. From the classification metrics (Table 2.39), the confusion matrix (Fig. 2.44), and the receiver operating characteristic (ROC) curves (Fig. 2.45), it is evident that classification of successive grades is challenging, and in particular classification metrics for grade 1 have low values in comparison to the other grades.

Figure 2.46 shows some examples of misclassification: grade 1 knee joints predicted as grade 0, 2, and 3. Figure 2.47 shows the misclassification of knee joints categorised as grade 0, 2 and 3 predicted as grade 1. These images show minimal variations in terms of joint space width and osteophytes formation, making them challenging to distinguish. Even the more serious misclassification in Fig. 2.48, for instance grade 0 predicted as grade 3 and vice versa, do not show very distinguishable variations. Furthermore, when the knee X-ray images belonging to grade 0 and grade 1 severity are examined, it can be seen that there are very subtle variations in

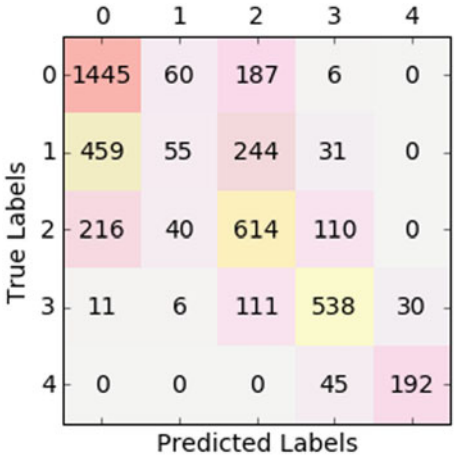


Fig. 2.44 Confusion matrix for the multi-class classification using the jointly trained CNN

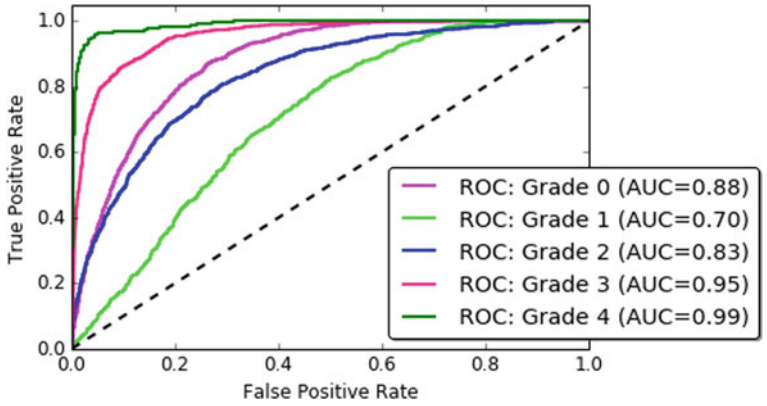


Fig. 2.45 ROC for the multi-class classification using the jointly trained CNN

terms of the joint space width and osteophytes formation. Even better representations are needed to capture these fine-grained variations and to distinguish coarse grades: grade 0 and grade 1 images.

Discussion. Jointly training a CNN from scratch using the multi-objective convolutional approach improves the multi-class classification accuracy and minimises the mean-squared error. However, successive grade classification still remains a challenge. Even though the KL grades are widely used for assessing knee OA severity in clinical settings, there has been continued investigation and criticism over the use of KL grades as the individual categories are not equidistant from each other [22, 24, 25, 81, 82]. This could be a reason for the low multi-class classification accuracy in the automatic quantification. Using OARSI readings instead of KL grades could

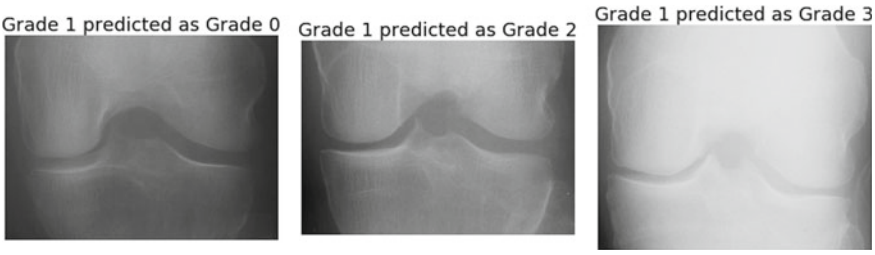


Fig. 2.46 Mis-classifications: grade 1 joints predicted as grade 0, 2, and 3

Fig. 2.47 Misclassification: other grade knee joints predicted as grade 1

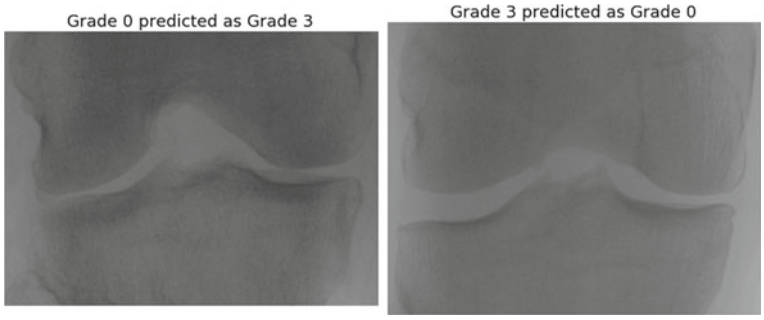


Fig. 2.48 An instance of more severe misclassification: grade 0 and grade 3

possibly provide better results for automatic quantification as the knee OA features such as joint space narrowing, osteophytes formation, and sclerosis are separately graded. Moreover, when the knee X-ray images belonging to grade 0 and grade 1 severity are visually examined, it can be seen that there are very subtle variations in terms of the joint space width and osteophytes formation. To capture these variations and distinguish these coarse grades, for instance grade 0 versus grade 1, even better representations are required. Indeed, it should also be recognised that even medical experts do not always agree upon a particular KL grade e.g. either 0 or 1 attributed to the initial stage of knee OA [22, 25, 81].

2.5.2.6 Ordinal Regression

Ordinal regression[8] is an intermediate task between multi-class classification and regression, sharing the properties of both. The outcomes or predictions in multi-class classification are discrete values and there is a meaningful order in the classes in regression. Ordinal regression is useful to classify patterns using a categorical scale which shows a natural order between the labels [83, 84]. The misclassification from a normal classifier are treated the same, that is no misclassification are worse than others [85]. However, some mis-classifications in ordinal regression, for instance the mis-classification on the extreme grades: grade 0 to grade 4 is treated worse than the others. This implies that the distances between the classes need to be taken into account when training a classifier. When quantifying the stages of a physical disease, it is preferable to predict the stage as 'mild' or 'doubtful' than 'absent' when the true label is 'severe'. Ordinal regression models formalise this notion of order by ensuring that predictions farther from the true label incur a greater penalty than those closer to the true label [84]. The author believes that the KL grades prediction based on ordinal regression can further improve classification performance by reducing the margin of error (mean-squared error), considering the progressive nature of knee OA and the ground truth or labels for training a CNN i.e. the KL grades in an ordinal scale (0–4).

CNN Configuration for Ordinal Regression. For ordinal regression output, the last stage of the CNN (Table 2.38) that gave best results on the joint training for multi-class classification and regression is modified. The previous approach on the joint training used two dense layers with softmax and linear activations in parallel (Fig. 2.41) for simultaneous multi-class classification and regression outputs. To train the CNN for ordinal regression, fixed weights ($[w_0, w_1, w_2, w_3, w_4] = [0, 1, 2, 3, 4]$) are applied to the outputs (probabilities) from the dense layer (Clsf) with softmax activations and back-propagate through a dense layer (Reg) with linear activations, optimising the mean-squared error loss function. The dense layer with softmax activations is treated as a hidden layer in this configuration. This is similar to the approach proposed by Beckham et al. [85] for ordinal classification. Figure 2.49 shows the CNN configuration for ordinal regression.

CNN Training. The CNN for ordinal regression (Table 2.41) is based on a lightweight architecture with ∼2.9 million free parameters in total and it contains eight layers with learned weights: seven convolutional layers and a fully connected layer. Each convolutional layer is followed by batch normalisation and a ReLU activation layer. To avoid overfitting, drop out with ratio 0.3 is applied after the last fully connected (fc5) layer. In addition to this, a L2 weight regularisation penalty of 0.01 is applied to all the convolutional and fully connected layers except the first two convolutional layers. The fc5 layer is followed by two dense layers with softmax (fc6-Clsf) and linear activations (fc7-Reg). The output of the softmax (fc6-Clsf) layer is multiplied (dot product) with fixed weights ([0, 1, 2, 3, 4]) and given as input to the last dense

[8] https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php.

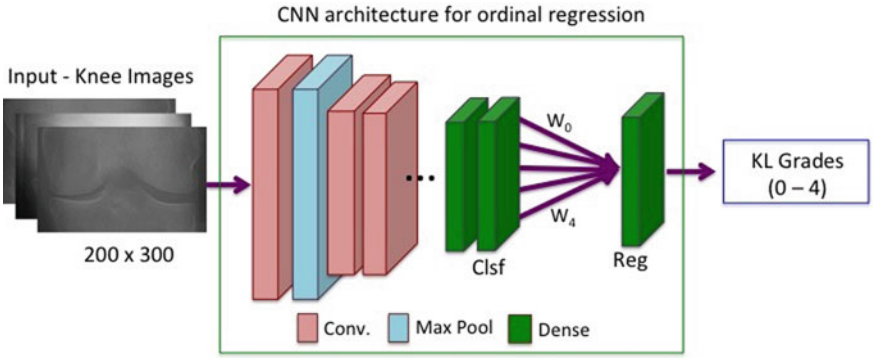


Fig. 2.49 The CNN configuration for ordinal regression

layer (fc7-Reg). This network is trained to minimise two loss functions: categorical cross-entropy and mean-squared error with equal weights. The dense layers (fc7-Reg) and (fc6-Clsf) provides the ordinal regression and multi-class classification outputs. The network is trained for 80 epochs with a batch size 32, using the Adam optimiser with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. The same training, validation, and test data are used from the joint training to make valid comparison of the results.

The CNN configuration in Table 2.41 gives the best results for ordinal regression and this configuration is similar to the network for joint training (Table 2.38) except the arrangement of the last two dense layers. Tuning the hyper-parameters of this CNN by increasing the number of layers with learned weights does not improve the quantification performance. Therefore, this CNN configuration is selected as the final network for ordinal regression.

Results. The learning curves (Fig. 2.50) obtained whilst training the CNN for ordinal regression shows convergence of the validation and training losses with improvement in validation and classification accuracies. Figure 2.51 shows the classification accuracy of the trained CNN model after every epoch on the test data for the ordinal regression and classification output. After 40 epochs of training, there is no significant improvement in the classification accuracies. There is a slight decrease in the accuracy of the ordinal regression in comparison to the classification. This is likely due to the rounding of the output.

After training, the CNN gives a classification accuracy of 64.3% on the test data. In the previous method on jointly training a CNN for classification and regression, a multi-class classification accuracy of 64.6% (Table 2.39) is achieved. As the same CNN configuration is used except the last stage (Fig. 2.49) and other settings are retained, the classification performance remains almost the same for the CNN trained for ordinal regression in comparison to the jointly trained CNN.

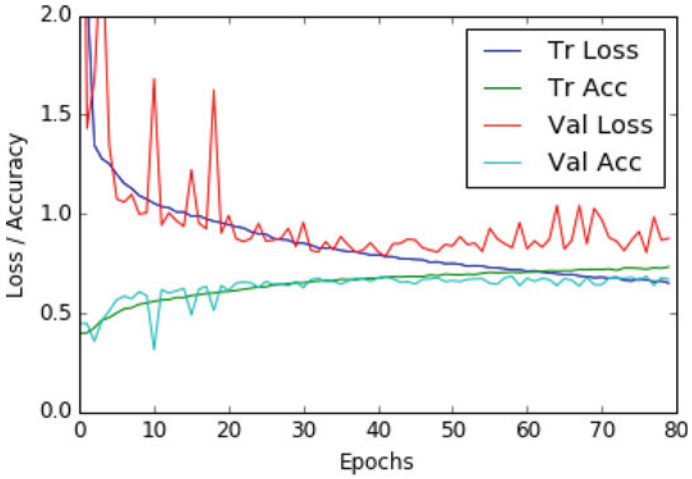
The classification metrics for the ordinal regression output are computed by rounding the predictions to integer values (0, 1, 2, 3, or 4). After rounding, the classification

Table 2.41 CNN architecture for ordinal regression

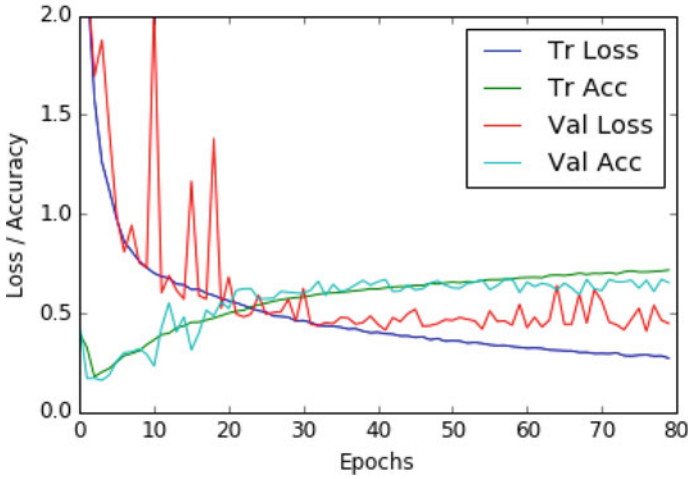
Layer	Kernels	Kernel size	Strides	Output shape
Input	–	–	–	$1 \times 200 \times 300$
conv1	32	11×11	2	$32 \times 100 \times 150$
maxPool1	–	3×3	2	$32 \times 49 \times 74$
conv2-1	64	3×3	1	$64 \times 49 \times 74$
conv2-2	64	3×3	1	$64 \times 49 \times 74$
maxPool2	–	3×3	2	$64 \times 24 \times 36$
conv3-1	96	3×3	1	$96 \times 24 \times 36$
conv3-2	96	3×3	1	$96 \times 24 \times 36$
maxPool3	–	3×3	2	$96 \times 11 \times 17$
conv4-1	128	3×3	1	$128 \times 11 \times 17$
conv4-2	128	3×3	1	$128 \times 11 \times 17$
maxPool4	–	3×3	2	$128 \times 5 \times 8$
fc5	–	–	–	512
fc6-Clsf	–	–	–	5
Input-weights	–	–	–	5
Merge-product	–	–	–	1
fc7-Reg	–	–	–	1

accuracy for the ordinal regression output is 61.8% with mean-squared error 0.504 on the test data. The classification metrics for the regression output from the jointly trained CNN gives a classification accuracy 53.3% with mean squared error 0.595 on the test data. There is an improvement in the classification accuracy and mean-squared error for ordinal regression in comparison to the previous regression results (see Table 2.40). Table 2.42 shows the precision, recall, and F_1 score for regression and ordinal regression. From the results it is evident that ordinal regression is out performing regression for quantifying knee OA images in a continuous scale.

Results Comparison. Four approaches are investigated to automatically quantify knee OA severity. Table 2.43 shows the multi-class classification accuracy and mean-squared error for the four approaches: (1) fine-tuning off-the-shelf CNN (BVLC CaffeNet), (2) training CNNs from scratch individually for classification and regression, (3) jointly training a CNN based on multi-objective convolutional learning for classification and regression, and (4) training a CNN for ordinal regression. These results are compared to WNDCHRM, that gave the previous best results for automatically classifying knee OA radiographs. The results show that jointly trained CNN gives the best results for multi-class classification. The ordinal regression outperforms all



(a) Classification



(b) Ordinal regression

Fig. 2.50 Learning curves for **a** classification and **b** ordinal regression

the other methods for quantifying knee OA images in a continuous scale with low mean-squared error.

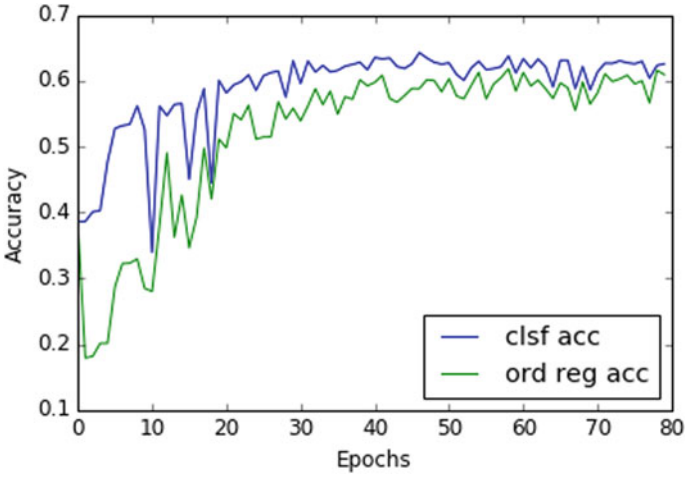


Fig. 2.51 The CNN configuration for ordinal regression

Table 2.42 Comparison of classification metrics from regression and ordinal regression

Grades	Regression			Ordinal regression		
	Precision	Recall	F_1	Precision	Recall	F_1
0	0.78	0.58	0.66	0.71	0.79	0.75
1	0.29	0.55	0.38	0.31	0.33	0.32
2	0.51	0.50	0.50	0.59	0.44	0.50
3	0.65	0.53	0.58	0.74	0.73	0.73
4	0.63	0.33	0.43	0.81	0.76	0.78
Mean	0.60	0.53	0.55	0.62	0.62	0.61

Table 2.43 Comparison of classification, regression and ordinal regression results

Method	Classification accuracy (%)	Mean-squared error
WNDCHRM	34.8	2.112
Fine-tuned BVLC CaffeNet	57.6	0.836
CNN-classification	61.8	0.735
CNN-regression	54.7	0.574
Jointly trained CNN	**64.6**	0.507
Ordinal regression	64.3	**0.480**

2.5.3 An Automatic Knee OA Diagnostic System

An automatic knee OA diagnostic system is developed combining the automatic localisation pipeline and the quantification pipeline developed in the previous sections. Figure 2.52 shows the proposed end-to-end diagnostic pipeline to automatically quantify knee OA severity based on KL grades. The input X-ray images are subjected to histogram equalisation, mean normalisation and resized to a fixed size 256×256. A fully convolutional network (FCN) is used to automatically detect the ROI, the knee joint regions. The bounding box coordinates of the ROI are calculated using simple contour detection. The knee joint regions are extracted from the knee radiographs using the bounding box coordinates. The localised and extracted knee images are resized to 200×300 to preserve the mean aspect ratio ($\sim$1.6) and fed to the jointly trained CNN. This system provides quantification of knee OA severity in both discrete grades (classification) and continuous grades (regression) simultaneously.

The major pathological features that indicate the onset of knee OA include: reduction in joint space width due to loss of knee cartilage, and the formation of bone spurs (osteophytes) or bony projections along the joint margins. The author believes that quantifying these features along with the KL grades can provide deeper insights to assess knee OA severity and to study the progression of knee OA. Therefore, a deep learning-based automatic knee OA diagnostic system that can provide simultaneous predictions of KL grades, JSN, and osteophytes is developed.

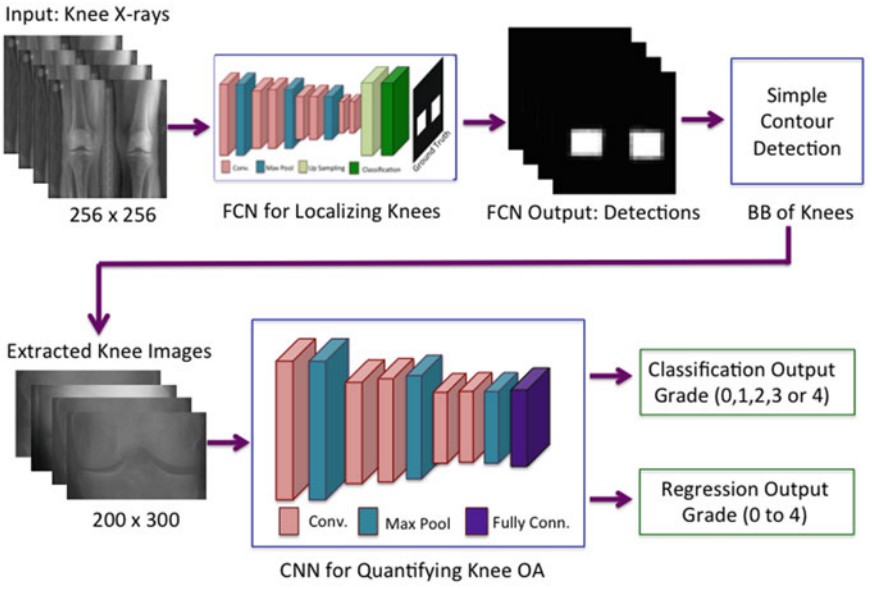


Fig. 2.52 The proposed pipeline for quantifying knee OA severity

2.5.4 Summary and Discussion

Four approaches are presented in this section to automatically assess knee OA severity using CNNs. First, the existing pre-trained CNNs are investigated for classifying knee images based on KL grades. Two methods are used in this approach: using the pre-trained CNNs for fixed feature extraction, and fine-tuning the pre-trained CNNs using the transfer learning approach. The predictions or outputs from these methods are ordinal KL grades (0, 1, 2, 3 or 4). Furthermore, the author argued that quantifying knee OA severity in a continuous scale (0–4) is more appropriate as the OA degradation is progressive in nature, not discrete. Regression is used to quantify the knee OA severity on a continuous scale. The classification and regression results from the proposed methods in this chapter outperform the previous best results achieved by WNDCHRM, which uses many hand-crafted features with a variation of k-nearest neighbour classifier for classifying knee OA radiographs.

Second, CNNs are trained from scratch for classification and regression. The objective was to further improve the quantification results. As the training data is relatively scarce, a lightweight architectures with fewer ($\sim$4 to $\sim$5 million) free parameters are used in the CNNs. The fully trained CNN for classification achieved high classification accuracy in comparison to the pre-trained CNNs. However, the fully trained CNN for regression did not achieve high-performing results as no ground truth of KL grades was available on a continuous scale. Therefore, the discrete KL grades are used to train the CNNs for regression.

Third, CNNs are fully trained using multi-objective learning for simultaneous classification and regression. The intuition behind this is optimising a CNN with two loss functions provide a stronger error signal and it is a step to improve the overall quantification, considering both classification and regression results. The jointly trained CNN achieved better quantification results with a high classification accuracy in comparison to the previous methods.

As a last approach, CNNs are fully trained for ordinal regression using a softmax dense layer as the hidden layer. This approach achieved low mean-squared error and outperformed other methods to quantify knee OA severity in a continuous scale. The added benefit of this method is to provide simultaneous multi-class classification output.

In summary, a progressive improvement is achieved in the quantification performance with an increase in classification accuracy and other performance metrics in the four approaches to automatically quantify knee OA severity. To conclude this Section, an error analysis was presented that discusses the possible reasons for the misclassification from the jointly trained CNN. The variations in the X-ray imaging protocols and discrepancies in the KL grades scoring need to be taken into account when analysing the mis-classification.

2.6 Conclusion

The main goal of this research is to advance the state-of-the-art in computer aided diagnostics of the severity of knee OA by developing deep learning based automatic methods. According to the literature, automatic assessment of knee OA severity has been previously approached as an image classification problem and existing approaches report low accuracy for multi-class and classification of successive grades. The state-of-the-art machine learning based methods are investigated for image classification, and we developed new methods using convolutional neural networks (CNNs) to automatically classify knee OA images. A significant outcome of this research is a new automatic knee OA diagnostic system that achieves high accuracy, on par with radiologic reliability readings, which are considered the gold standard for knee OA assessment.

In recent years, deep learning-based approaches, in particular CNNs, have become highly successful in many computer vision tasks and medical applications. Our research has mainly focused on developing a deep learning-based computer aided diagnostic system. The proposed approaches in this chapter are related to two main medical applications: localising or automatically detecting and extracting a region of interest (ROI) from a radiograph, and classifying the ROI to automatically assess disease severity. The FCN-based localisation approach could be extended to other medical applications such as localising a substructure or a ROI in MRI and CT scan images, object or lesion detection, locating anatomical landmarks or identifying imaging markers to study the disease progression. For instance, we followed a similar FCN-based approach to automatically detect and quantify ischemic injury (brain lesions) on diffusion-weighted MRI of infants, and we improved the state-of-the-art by achieving promising results.

In the author's opinion the most interesting research findings in this research are as follows. First, fine-tuning off-the-shelf CNNs pre-trained on very large datasets such as ImageNet (with $\sim$1M images) to classify knee images with relatively small datasets (with $\sim$10,000 images) is promising for medical image classification. The main challenge in medical image classification is a lack of sufficient annotated data for training deep networks from scratch. Fine-tuning existing CNNs that have been trained using a large annotated dataset from a different application is possibly the best alternative to full training for medical applications. A second extremely interesting result is that training CNNs, optimising a weighted ratio of two loss functions for simultaneous classification and regression, provides a better error signal to the network and improves the overall classification performance. Many diseases are progressive by nature such as Alzheimer's disease, cancer, emphysema, tumours, lesions, and muscular dystrophy. Automatic quantification of such diseases using jointly trained CNNs may improve the quantification performance and provide insights to study the progression of the disease. Finally, it is very interesting that using multi-objective convolutional learning to jointly train CNNs based on different diagnostic features of a disease as the ground truth, can produce an overall improvement in the quantification performance achieving, results on par with human accuracy. Multi-objective

learning and joint prediction of multiple regression and classification variables may be useful to assess diseases involving multiple diagnostic features like Alzheimer's, multiple sclerosis, and multiple myeloma (cancer).

There are several potential directions for future work and further development of the research in this chapter. Some of the interesting extensions and prospects are outlined as follows.

Training an end-to-end deep learning model: The knee OA diagnostic pipeline consists of two steps: (1) localising the knee joints in radiographs and (2) assessing the knee OA severity from the localised knee joints. A FCN was trained for automatic localisation and a CNN was jointly trained for classification and regression of knee OA images. It would be interesting to train a single deep learning model integrating the FCN for localisation and the CNN for classification and/or regression, as this could further improve the automatic assessment of knee OA. In a recent work, Górriz et al. [86] presented a CNN attention-based end-to-end architecture to automatically assess knee OA severity.

Using semantic segmentations to measure joint space width: Among the knee OA diagnostic features, joint space narrowing is highly sensitive to changes due to disease progression. The proposed approach to automatically localise the knee joints using fully convolutional network can be extended for semantic segmentation of the knee joints and can be used to automatically measure the joint space width (JSW) between the femur and tibia. However, pixel level knee joint annotations in radiographs are needed to measure the JSW.

Assessing the progression of knee OA severity: The automatic quantification methods developed in this study can be extended to assess the progression and early detection of knee OA severity. The baseline datasets are used from the OAI and the MOST dataset. Datasets are available for annual follow-up visits up to 9 years. These datasets could be used to detect the features predictive of radiographic knee OA progression. Shamir et al. reported a similar approach using WNDCHRM that predicted whether a knee would change from KL grade 0 to grade 3 with 72% accuracy using 20 years of data [3]. The findings of our current work has indicated the ability to improve upon the existing methods.

Relating the automatic quantification results to knee pain: The primary clinical features to assess knee OA are pain and radiographic evidence of deformity [87]. It would be interesting to study the relationship between the automatic assessments of the proposed methods (KL grades) to WOMAC scores for knee pain. WOMAC is one among the most widely used assessments in knee OA.

Relating the automatic quantification results to physiological variables: There are several pathological and physiological variables available in the OAI and the MOST datasets. These variables include potential predictors of knee pain status. It would be interesting to study the relationship between the outcomes of the automatic methods and the predictions from the pathological and physiological variables. Abedin et al. [88] presented a comparative analysis to predict knee OA severity based on statistical models developed on physiological variables and a CNN model developed on features from X-ray images.

Investigating human level accuracy: The radiologic reliability readings from the OAI used 150 participants (300 knees) to evaluate the test-retest reliability of semi-quantitative readings. This is considered the current gold standard for knee OA assessment. Simple kappa coefficients and weighted kappa coefficients were used to evaluate the inter-rater agreement. Investigating the human level accuracy for a large sample and comparing it with the automatic quantification results would provide insight to help reduce the error involved in automatic assessments.

Acknowledgements This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant numbers SFI/12/RC/2289 and 15/SIRG/3283.
The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health.
MOST is comprised of four cooperative grants (Felson—AG18820; Torner—AG18832; Lewis—AG18947; and Nevitt—AG19069) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by MOST study investigators. This manuscript was prepared using MOST data and does not necessarily reflect the opinions or views of MOST investigators.

References

1. Shamir, L., Ling, S.M., Scott Jr., W.W., Bos, A., Orlov, N., Macura, T.J., Eckley, D.M., Ferrucci, L., Goldberg, I.G.: Knee X-ray image analysis method for automated detection of osteoarthritis. IEEE Trans. Biomed. Eng. **56**(2) (2009)
2. Thomson, J., ONeill, T., Felson, D., Cootes, T.: 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, pp. 127–134. Springer (2015)
3. Shamir, L., Ling, S.M., Scott, W., Hochberg, M., Ferrucci, L., Goldberg, I.G.: Early detection of radiographic knee osteoarthritis using computer-aided analysis. Osteoarthr. Cartil. **17**(10), 1307–1312 (2009)
4. Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J., Goldberg, I.G.: WNDCHARM—an open source utility for biological image analysis. Source Code Biol. Med. **3**(1), 13 (2008)
5. Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., Goldberg, I.G.: WND-CHARM: multi-purpose image classification using compound image transforms. Pattern Recognit. Lett. **29**(11), 1684–1693 (2008)
6. Oka, H., Muraki, S., Akune, T., Mabuchi, A., Suzuki, T., Yoshida, H., Yamamoto, S., Nakamura, K., Yoshimura, N., Kawaguchi, H.: Fully automatic quantification of knee osteoarthritis severity on plain radiographs. Osteoarthr. Cartil. **16**(11), 1300–1306 (2008)
7. Park, H.J., Kim, S.S., Lee, S.Y., Park, N.H., Park, J.Y., Choi, Y.J., Jeon, H.J.: A practical MRI grading system for osteoarthritis of the knee: association with Kellgren-Lawrence radiographic scores. Eur. J. Radiol. **82**(1), 112–117 (2013)
8. Lee, H.: Unsupervised feature learning via sparse hierarchical representations. Ph.D. thesis, Stanford University (2010)
9. Le, Q.V.: Scalable feature learning. Ph.D. thesis, Stanford University (2013)
10. Yang, S.: Feature engineering in fine-grained image classification. Ph.D. thesis, University of Washington (2013)

11. Donoghue, C.R.: Analysis of MRI for knee osteoarthritis using machine learning. Ph.D. thesis, Imperial College London (2013)
12. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Osteoarthr. Cartil. **42**(1), 60–88 (2017)
13. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. (0) (2017)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
16. Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M.: Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: International Conference on Medical Image Computing and Computer-assisted Intervention, 246–253. Springer (2013)
17. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
18. Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage **108**, 214–224 (2015)
19. Roth, H.R., Farag, A., Lu, L., Turkbey, E.B., Summers, R.M.: Deep convolutional networks for pancreas segmentation in CT imaging. In: SPIE Medical Imaging, International Society for Optics and Photonics, pp. 94131G–94131G (2015)
20. Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2843–2851 (2012)
21. Marijnissen, A.C., Vincken, K.L., Vos, P.A., Saris, D., Viergever, M., Bijlsma, J., Bartels, L., Lafeber, F.: Knee images digital analysis (kida): a novel method to quantify individual radiographic features of knee osteoarthritis in detail. Osteoarthr. Cartil. **16**(2), 234–243 (2008)
22. Felson, D.T., McAlindon, T.E., Anderson, J.J., Weissman, B.W., Aliabadi, P., Evans, S., Levy, D., LaValley, M.P.: Defining radiographic osteoarthritis for the whole knee. Osteoarthr. Cartil. **5**(4), 241–250 (1997)
23. Braun, H.J., Gold, G.E.: Diagnosis of osteoarthritis: imaging. Bone **51**(2), 278–288 (2012)
24. Emrani, P.S., Katz, J.N., Kessler, C.L., Reichmann, W.M., Wright, E.A., McAlindon, T.E., Losina, E.: Joint space narrowing and kellgren-lawrence progression in knee osteoarthritis: an analytic literature synthesis. Osteoarthr. Cartil. **16**(8), 873–882 (2008)
25. Hart, D., Spector, T.: Kellgren & Lawrence grade 1 osteophytes in the kneedoubtful or definite? Osteoarthr. Cartil. **11**(2), 149–150 (2003)
26. Shaikh, H., Panbude, J., Joshi, A.: Image segmentation techniques and its applications for knee joints: a survey. IOSR J. Electron. Commun. Eng. (IOSR-JECE) **9**(5), 23–28 (2014)
27. Sun, Y., Teo, E., Zhang, Q.: Discussions of knee joint segmentation. In: International Conference on Biomedical and Pharmaceutical Engineering, 2006. ICBPE 2006. IEEE (2006)
28. Gornale, S.S., Patravali, P.U., Manza, R.R.: Detection of osteoarthritis using knee x-ray image analyses: a machine vision based approach. Int. J. Comput. Appl. **145**(1) (2016)
29. Tiulpin, A., Thevenot, J., Rahtu, E., Saarakkala, S.: A novel method for automatic localization of joint area on knee plain radiographs. In: Scandinavian Conference on Image Analysis, pp. 290–301. Springer (2017)
30. Stammberger, T., Eckstein, F., Michaelis, M., Englmeier, K.H., Reiser, M.: Interobserver reproducibility of quantitative cartilage measurements: comparison of b-spline snakes and manual segmentation. Magn. Reson. Imaging **17**(7), 1033–1042 (1999)

31. Cohen, Z.A., Mccarthy, D.M., Kwak, S.D., Legrand, P., Fogarasi, F., Ciaccio, E.J., Ateshian, G.A.: Knee cartilage topography, thickness, and contact areas from mri: in-vitro calibration and in-vivo measurements. Osteoarthr. Cartil. **7**(1), 95–109 (1999)
32. Hirvasniemi, J., Thevenot, J., Immonen, V., Liikavainio, T., Pulkkinen, P., Jämsä, T., Arokoski, J., Saarakkala, S.: Quantification of differences in bone texture from plain radiographs in knees with and without osteoarthritis. Osteoarthr. Cartil. **22**(10), 1724–1731 (2014)
33. Woloszynski, T., Podsiadlo, P., Stachowiak, G., Kurzynski, M.: A signature dissimilarity measure for trabecular bone texture in knee radiographs. Med. Phys. **37**(5), 2030–2042 (2010)
34. Zhao, F., Xie, X.: An overview of interactive medical image segmentation. Ann. BMVA **2013**(7), 1–22 (2013)
35. Pirnog, C.D.: Articular cartilage segmentation and tracking in sequential MR images of the knee. Ph.D. thesis, ETH Zurich (2005)
36. Duryea, J., Li, J., Peterfy, C., Gordon, C., Genant, H.: Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. Med. Phys. **27**(3), 580–591 (2000)
37. Podsiadlo, P., Wolski, M., Stachowiak, G.: Automated selection of trabecular bone regions in knee radiographs. Med. Phys. **35**(5), 1870–1883 (2008)
38. Anifah, L., Purnama, I.K.E., Hariadi, M., Purnomo, M.H.: Automatic segmentation of impaired joint space area for osteoarthritis knee on x-ray image using gabor filter based morphology process. IPTEK J. Technol. Sci. **22**(3) (2011)
39. Lee, H.C., Lee, J.S., Lin, M.C.J., Wu, C.H., Sun, Y.N.: Automatic assessment of knee osteoarthritis parameters from two-dimensional x-ray image. In: First International Conference on Innovative Computing, Information and Control, 2006. ICICIC'06, vol. 2, pp. 673–676. IEEE (2006)
40. Subramoniam, M., Rajini, V.: Local binary pattern approach to the classification of osteoarthritis in knee x-ray images. Asian J. Sci. Res. **6**(4), 805 (2013)
41. Subramoniam, B., et al.: A non-invasive computer aided diagnosis of osteoarthritis from digital x-ray images. Biomed. Res. (2015)
42. Deokar, D.D., Patil, C.G.: Effective feature extraction based automatic knee osteoarthritis detection and classification using neural network. Int. J. Eng. Tech. **1**(3) (2015)
43. Yoo, T.K., Kim, D.W., Choi, S.B., Park, J.S.: Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PLoS One **11**(2), e0148724 (2016)
44. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci. Rep. **8**(1), 1727 (2018)
45. Antony, J., McGuinness, K., O'Connor, N.E., Moran, K.: Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1195–1200. IEEE (2016)
46. Sobel, I.: An isotropic 3×3 image gradient operator. In: Machine Vision for Three-dimensional Sciences (1990)
47. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 1, pp. 886–893. CVPR 2005. IEEE (2005)
48. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
49. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
50. Kayalibay, B., Jensen, G., van der Smagt, P.: CNN-based segmentation of medical imaging data. arXiv:1701.03056 (2017)
51. Christ, P.F., Ettlinger, F., Grün, F., Elshaera, M.E.A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P., et al.: Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. arXiv:1702.05970 (2017)

52. Li, F.F., Karpathy, A., Johnson, J.: CS231n: convolutional neural networks for visual recognition. http://cs231n.github.io/convolutional-networks/ (2016)
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
54. Kilian, J.: Simple image analysis by moments. OpenCV library documentation (2001)
55. Bressan, M., Dance, C.R., Poirier, H., Arregui, D.: Local contrast enhancement. In: Color Imaging: Processing, Hardcopy, and Applications, p. 64930Y (2007)
56. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process **39**(3), 355–368 (1987)
57. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2147–2154 (2014)
58. Antony, J., McGuinness, K., Moran, K., OConnor, N.E.: 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, pp. 376–390. Springer (2017)
59. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
60. Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J., Goldberg, I.: Wnd-charm: multipurpose image classifier. Astrophysics Source Code Library (2013)
61. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Syst. Man Cybern. **8**(6), 460–473 (1978)
62. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973)
63. Grigorescu, S.E., Petkov, N., Kruizinga, P.: Comparison of texture features based on gabor filters. IEEE Trans. Image Proc. **11**(10), 1160–1167 (2002)
64. Teague, M.R.: Image analysis via the general theory of moments. JOSA **70**(8), 920–930 (1980)
65. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of British Machine Vision Conference (2014)
66. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678 (2014)
67. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. In: Proceedings of the British Machine Vision Conference. BMVA Press (2014)
68. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
69. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)
70. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
71. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
72. Miao, S., Wang, Z.J., Zheng, Y., Liao, R.: Real-time 2d/3d registration via CNN regression. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1430–1434. IEEE (2016)
73. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in x-ray images. Med. Image Anal. **36**, 41–51 (2017)

74. Roth, H.R., Wang, Y., Yao, J., Lu, L., Burns, J.E., Summers, R.M.: Deep convolutional networks for automated detection of posterior-element fractures on spine CT. In: Proceedings Volume 9785, Medical Imaging 2016: Computer-Aided Diagnosis, SPIE Medical Imaging (2016)
75. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
76. Liu, S., Yang, J., Huang, C., Yang, M.H.: Multi-objective convolutional learning for face labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3451–3459 (2015)
77. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
78. Ranftl, R., Pock, T.: A deep variational model for image segmentation. In: German Conference on Pattern Recognition, pp. 107–118. Springer (2014)
79. Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E.: Toward automatic phenotyping of developing embryos from videos. IEEE Trans. Image Proc. **14**(9), 1360–1371 (2005)
80. Rudd, E.M., Günther, M., Boult, T.E.: Moon: a mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision, pp. 19–35. Springer (2016)
81. Schiphof, D., Boers, M., Bierma-Zeinstra, S.M.: Differences in descriptions of kellgren and lawrence grades of knee osteoarthritis. Ann. Rheum. Dis. **67**(7), 1034–1036 (2008)
82. Shamir, L., Felson, D.T., Ferrucci, L., Goldberg, I.G.: Assessment of osteoarthritis initiative-kellgren and lawrence scoring projects quality using computer analysis. J. Musculoskelet. Res. **13**(04), 197–201 (2010)
83. Gutiérrez, P.A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., Hervas-Martinez, C.: Ordinal regression methods: survey and experimental study. IEEE Trans. Knowl. Data Eng. **28**(1), 127–146 (2016)
84. Pedregosa, F., Bach, F., Gramfort, A.: On the consistency of ordinal regression methods. J. Mach. Learn. Res. **18**(55), 1–35 (2017)
85. Beckham, C., Pal, C.: A simple squared-error reformulation for ordinal classification. arXiv:1612.00775 (2016)
86. Górriz, M., Antony, J., McGuinness, K., Giró-i Nieto, X., OConnor, N.E.: Assessing knee OA severity with CNN attention-based end-to-end architectures. In: International Conference on Medical Imaging with Deep Learning, pp. 197–214 (2019)
87. Williams, D.A., Farrell, M.J., Cunningham, J., Gracely, R.H., Ambrose, K., Cupps, T., Mohan, N., Clauw, D.J.: Knee pain and radiographic osteoarthritis interact in the prediction of levels of self-reported disability. Arthritis Care Res. **51**(4), 558–561 (2004)
88. Abedin, J., Antony, J., McGuinness, K., Moran, K., OConnor, N.E., Rebholz-Schuhmann, D., Newell, J.: Predicting knee osteoarthritis severity: comparative modeling based on patients data and plain x-ray images. Sci. Rep. **9**(1), 5761 (2019)

Chapter 3
Classification of Tissue Regions in Histopathological Images: Comparison Between Pre-trained Convolutional Neural Networks and Local Binary Patterns Variants

Jakob N. Kather, Raquel Bello-Cerezo, Francesco Di Maria, Gabi W. van Pelt, Wilma E. Mesker, Niels Halama and Francesco Bianconi

Abstract The identification of tissue regions within histopathological images represents a fundamental step for diagnosis, patient stratification and follow-up. However, the huge amount of image data made available by the ever improving whole-slide imaging devices gives rise to a bottleneck in manual, microscopy-based evaluation. Furthermore, manual procedures generally show a significant intra- and/or inter-observer variability. In this scenario the objective of this chapter is to investigate the effectiveness of image features from last-generation, pre-trained convolutional net-

J. N. Kather
Department of Medicine III, University Hospital RWTH, Pauwelsstraße 30, 52074 Aachen, Germany
e-mail: jakob.kather@nct-heidelberg.de

R. Bello-Cerezo · F. Di Maria · F. Bianconi (✉)
Department of Engineering, Università degli studi di Perugia, Via Goffredo Duranti 93, 06125 Perugia, Italy
e-mail: bianco@ieee.org

R. Bello-Cerezo
e-mail: bellocerezo@gmail.com

F. Di Maria
e-mail: francesco.dimaria@unipg.it

N. Halama
National Center for Tumour Diseases, Universität Heidelberg, Neuenheimer Feld 460, 69120 Heidelberg, Germany
e-mail: niels.halama@nct-heidelberg.de

G. W. van Pelt · W. E. Mesker
Department of Surgery, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
e-mail: G.W.van_Pelt@lumc.nl

W. E. Mesker
e-mail: W.E.Mesker@lumc.nl

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_3

works against variants of Local Binary Patterns for classifying tissue sub-regions into meaningful classes such as epithelium, stroma, lymphocytes and necrosis. Experimenting with seven datasets of histopathological images we show that both classes of methods can be quite effective for the task, but with a noticeable superiority of descriptors based on convolutional neural networks. In particular, we show that these can be seamlessly integrated with standard classifiers (e.g. Support Vector Machines) to attain overall discrimination accuracy between 95 and 99%.

3.1 Introduction

It is generally accepted that the relative amount of different tissue structures such as epithelium, stroma, lymphocytes and necrosis plays a crucial role in a number of neoplastic disorders. In the last few years the tumour-stroma ratio (TSR) has for instance emerged as an important prognostic indicator in a wide number of oncologic diseases including colorectal cancer [1–3], epithelial ovarian cancer [4], hepatocellular carcinoma [5], nasopharyngeal cancer [6] and some forms of breast cancer [7–9]. Likewise, lymphocyte infiltration has been recognised as a prognostic factor in breast [10], colorectal [11–13] and prostate cancer [14]; whereas necrosis has been investigated as a potentially meaningful bio-marker in renal cell, lung, thyroid and colorectal carcinoma [15].

Pathologists routinely examine tissue slides and score them on the basis of those tumour micro-environmental features that are expected to provide useful prognostic information [16]. However, as Linder et al. [17] correctly noted, the huge amount of image data made available by the ever improving whole-slide imaging devices gives rise to a bottleneck in manual, microscopy-based evaluation. Furthermore, manual procedures generally show a very pronounced intra- and/or inter-observer variability. Courrech-Staal et al. [18] for instance found inter-observer agreement at grading adenocarcinoma biopsies into two and four TSR classes respectively in the range 81–98% and 51–72%; whereas Fuchs and Buhmann [19] reported 58% and 69% respectively intra- and inter-observer agreement at classifying cell nuclei as normal or atypical on tissue micro-arrays (TMA) from clear renal cell carcinoma.

Computer-assisted analysis of digital slides through quantitative and repeatable image-processing methods is currently being investigated as a potential means for overcoming these difficulties. In recent years the advent of convolutional neural networks (CNN) has represented a major breakthrough in the field of computer vision, leading to major improvements in areas like object, scene and face recognition [20–22]. Convolutional networks are usually composed of a variable number of layers (the more the layers, the 'deeper' the network) each containing a set of parameters whose values are to be determined through suitable training procedures [23]. Such procedures usually require a huge number of labelled images—typically by the order of the hundreds of thousands or more [24]. Training a network from scratch can be therefore difficult in histology, where the image availability is likely to be limited by privacy restrictions as well as practical and ethical reasons. CNN also proved quite

robust to domain adaptation, which means that networks trained in a certain domain can be seamlessly and effectively used in different contexts [25, 26]. Based on these considerations this chapter investigates the effectiveness of pre-trained networks for application in histology—specifically for the classification of tissue sub-regions in histopathological images. Pre-trained CNN features are compared with a set of Local Binary Patterns (LBP) variants [27, 28]. Albeit intrinsically different, these two strategies can be considered equivalent from a practical standpoint in that they can both be used 'out-of-the-box' without any further adjustments. This a clear advantage when it comes to designing computer-assisted systems for tissue analysis.

We show that image features from last generation pre-trained CNN can be seamlessly coupled with standard classifiers (e.g. nearest neighbourhood and Support Vector Machines) to classify tissue sub-types with overall accuracy in the range 95–99% with respect to a ground truth based on visual inspection. Texture descriptors based on LBP variants also proved effective for the task, but their accuracy was on average slightly worse than was obtained with CNN-based features.

3.2 Background and Related Work

Meijer et al. define image analysis as a special discipline in pathology aiming at obtaining 'diagnostically important information in an objective and reproducible manner' [29]. During the last fifteen years pathology has profited from the quick improvement of image acquisition systems which culminated in the creation of digital slide scanners. These devices produce whole-slide images (WSI) and therefore combine the advantages of digital cameras (high resolution; easy filing, recovery and transmission) with those of live microscopy such as whole slide access [30]. Besides, digital WSI are amenable to being processed by standard image processing and machine learning methods. Early applications of image processing methods in clinical pathology can be traced back to the 80s and included morphometry and counting [29]. Currently, automated methods can be used for detecting and segmenting objects and regions in tasks such as segmentation of tissue and tissue components, detection and segmentation of nuclei, segmentation of tubules, assessment of proliferation, detection of mitotic figures, etc.; as well as for computer-aided diagnosis (CAD), patient stratification and prognosis. For a broader overview on the subject we refer the reader to Refs. [31–35].

Among the possible applications, automatic classification of tissue regions into different classes such as epithelium, stroma, lymphocytes, necrosis and the like has, for the reasons mentioned in Sect. 3.1, generated considerable research interest in recent years. In particular, image-processing approaches based on texture analysis proved very suitable for this task. No surprise, then, that texture analysis has recently received much attention in this context. Diamond et al. [36] for instance reported 79.3% accuracy using grey-level co-occurrence features for classifying tissue regions from prostatic neoplasia into stroma, epithelium and normal tissue. Linder et al. [17] obtained 96.8% accuracy using Local Binary Patterns (LBP) and a contrast measure for identifying tumour epithelium and stroma in tissue microarrays

from colorectal cancer. Later on this value was improved to 96.9% by Bianconi et al. [37] by means of features based on visual perception. On the same dataset Nava et al. [38] recently attained 96.5% accuracy through Discrete Tchebichef Moments. More recently, Kather et al. [39] carried out an extensive comparison of texture descriptors on a new multi-class dataset of histological images from colorectal cancer and reported 98.6% accuracy for two-class tissue classification (i.e.: epithelium vs. stroma) and 87.4% for multi-class classification (i.e.: epithelium, simple stroma, complex stroma, immune cells, debris, mucosal glands, adipose tissue and background). Of late, Badejo et al. [40] evaluated six grey-scale LBP variants against one pre-trained CNN (AlexNet) for a number of medical image classification tasks. Their results showed comparable accuracy, but a better performance of the CNN overall.

In recent years the advent of Convolutional Neural Networks (CNN) has marked a turning point in the field and brought significant changes to the whole scenario [41]. Deep learning methods have in fact been shown to achieve state-of-the art results in a number of tasks such as tissue classification [32, 42–44], cell classification [45], mitosis detection [46] as well as nuclei and tubule segmentation [32].

There are basically three strategies whereby CNN can be used in medical image analysis [32, 47]: *full training*, *fine tuning* and *transfer learning*.

In full training a convolutional network is trained from scratch. To this end one can either pick one of the many architectures available or design a new one. In either case all the free parameters of the network need to be determined by training. This is no easy task for at least three reasons: (a) the need for a large dataset of labelled data, (b) the extensive computational and memory resources required and (c) the convergence issues that usually occur in the training phase [41, 47]. As a consequence full training is rarely done in histology—though attempts have been made with very compact nets as for instance in [44]. Fine tuning consists of taking a pre-trained net and optimising some of its parameter to fit a domain of application different from the one the net has been trained for. This process usually starts from the last layers of the network and proceeds backwards; thus we can have 'shallow tuning', which imbues only few of the last layers, or 'deep tuning', where more layers are involved. This procedure still requires medium-size datasets of labelled images—which again are not easily available in the medical domain. Finally, transfer learning consists of using pre-trained networks as 'off-the-shelf' feature extractors in combination with standard classifiers (e.g.: nearest neighbour, SVM, random forest, etc.). This approach proved effective in tasks including pulmonary nodule classification in CT scans [48] and lesion classification in mammographies [49].

Put into perspective, we believe that transfer learning is the most appealing strategy—at least from a practical standpoint—for its being conceptually simple, computationally cheap and free from tedious and time-consuming training procedures. Furthermore, the availability of open-access libraries, which are continuously improved and enriched with increasingly effective and accurate models (e.g.: MatConvNet [50], Caffe [51]), is a significant point of strength of this approach. Based on these considerations this work investigates the effectiveness of this scheme for the classification of tissue sub-regions in histopathological images. For comparison purposes the study includes a set of hand-designed image descriptors of the LBP family, which in previous studies [17, 39] have proved to be effective for this task.

3.3 Materials

We based the experiments on seven datasets of histopathological images from different sources. All the dataset contain sets of labelled images each representing well-defined tissue sub-regions. For each dataset the ground truth was established by human experts through visual inspection. Five of the datasets include samples of tissue epithelium and stroma only; the other feature additional classes such as adipose tissue, debris, lymphocytes, necrosis, etc. Notably, datasets 'Two-class LUMC' and 'Multi-class LUMC' are presented for the first time in this work.

3.3.1 Two-Class Datasets

3.3.1.1 Epistroma

This dataset is based on tissue samples from a cohort of 643 patients with colorectal cancer who underwent surgery at Helsinki University Central Hospital (Finland) from 1989 to 1998 [17, 52]. The images were immunostained for epidermal growth factor receptor (EGFR) using diaminobenzidine (DAB) as a brown chromogen and hematoxylin as an unspecific blue counterstain. Image acquisition was based on a Zeiss Axioskop 2 MOT microscope (Zeiss Gmbh, Göttingen, Germany) and a CCD camera (Zeiss Axiocam HR) [53]. The samples were digitized at a fixed resolution of 0.26 μm/px and separated into well-defined regions representing either epithelium or stroma. The resulting dataset is composed of 1376 images (825 epithelium and 551 stroma) of dimension ranging from 93 to 2372 px in width and from 94 to 2373 px in height. Sample images are available in Fig. 3.1.

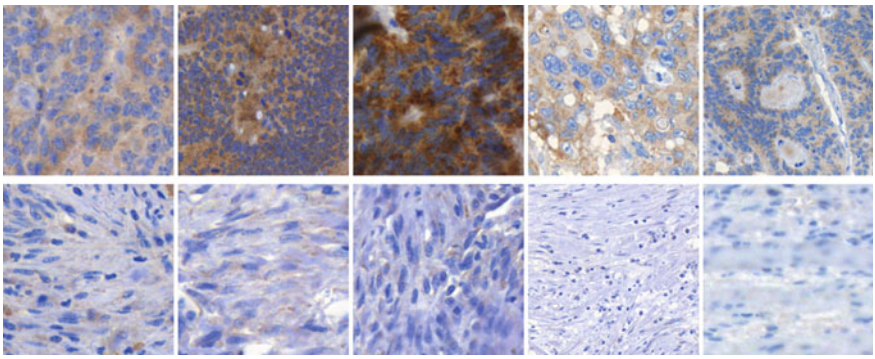


Fig. 3.1 Epistroma dataset. Samples of epithelium (first row) and stroma (second row)

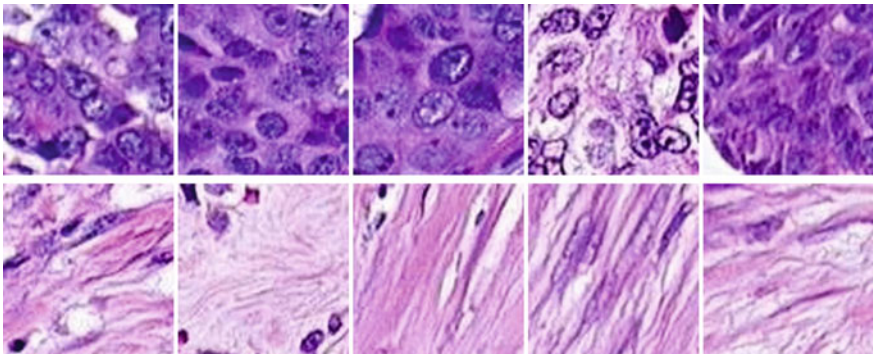


Fig. 3.2 NKI dataset. Samples of epithelium (first row) and stroma (second row)

3.3.1.2 NKI

This dataset features 1295 images representing well defined samples of either *epithelium* (1106 images) or *stroma* (189) from breast cancer. The original TMAs [54] come from a cohort of 248 patients (women younger than 53 years with stage I or II breast cancer) enrolled at Netherlands Cancer Institute (Amsterdam, The Netherlands). Details about the digitisation apparatus and scanning resolution are not available. The images come with a manually predefined ground-truth segmentation into epithelium and stroma [55] which we used to crop tiles of dimension 100 px × 100 px such as that at least 90% of the area of each tile represented either epithelium or stroma. We further screened the images for quality or differences in the evaluation compared with the one provided in [54]. Sample images are available in Fig. 3.2.

3.3.1.3 Two-Class Kather's

This is a reduced version of Multi-class Kather's dataset (Sect. 3.3.2.1) in which we only considered labelled samples of *epithelium* and *stroma* and discarded the other classes (Fig. 3.4).

3.3.1.4 Two-Class LUMC

Similarly to Two-class Kather's, this is a subset of Multi-class LUMC dataset (Sect. 3.3.2.2) in which we only experimented with samples of *epithelium* and *stroma* (Fig. 3.5).

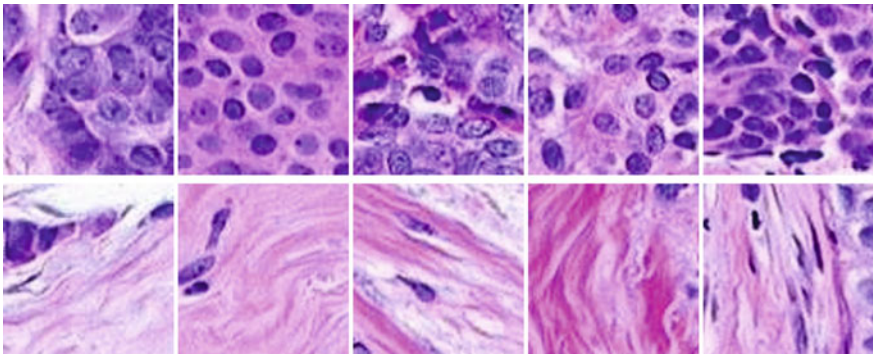


Fig. 3.3 VGH dataset. Samples of epithelium (first row) and stroma (second row)

3.3.1.5 VGH

Dataset VGH contains 273 images of *epithelium* (226 images) and *stroma* (47) from breast cancer. The original TMAs come from 328 patients enrolled at Vancouver General Hospital (Vancouver, Canada). Differently from the NKI dataset, in this case the cohort of patients included a higher proportion of older women and with more advanced stages of the disease [54]. As in NKI, no further details about the digitisation apparatus and scanning resolution are available. Following the same procedure used for NKI we cropped the original images into tiles of dimension 100 px $\times$ 100 px representing well-defined regions of either epithelium or stroma. We further screened the resulting images for quality and/or disagreement with the ground truth provided in [54]. Sample images are available in Fig. 3.3.

3.3.2 Multi-class Datasets

3.3.2.1 Multi-class Kather's

The recently released Kather's multi-class dataset [39] is a collection of 5000 image tiles, 625 for each of the following classes: *epithelium*, *simple stroma*, *complex stroma*, *immune cells*, *debris*, *normal mucosal glands*, *adipose tissue* and *background* (i.e.: no tissue). The images have a dimension of 150 px $\times$ 150 px. The samples come from ten H&E stained tissue slides of low- and high-grade colorectal cancer obtained at the University Medical Center Mannheim (Mannheim, Germany). The slides were digitised through an Aperio ScanScope (Aperio/Leica biosystems, Wetzlar, Germany) and saved as compressed Aperio `.svs` files [56]. The spatial resolution of the slides is approximately 0.5 µm/px, and the samples exhibit different overall brightness which reflects the variability that routinely affects histopathological slides. The dataset is available under Creative Common Attribution 4.0 Interna-

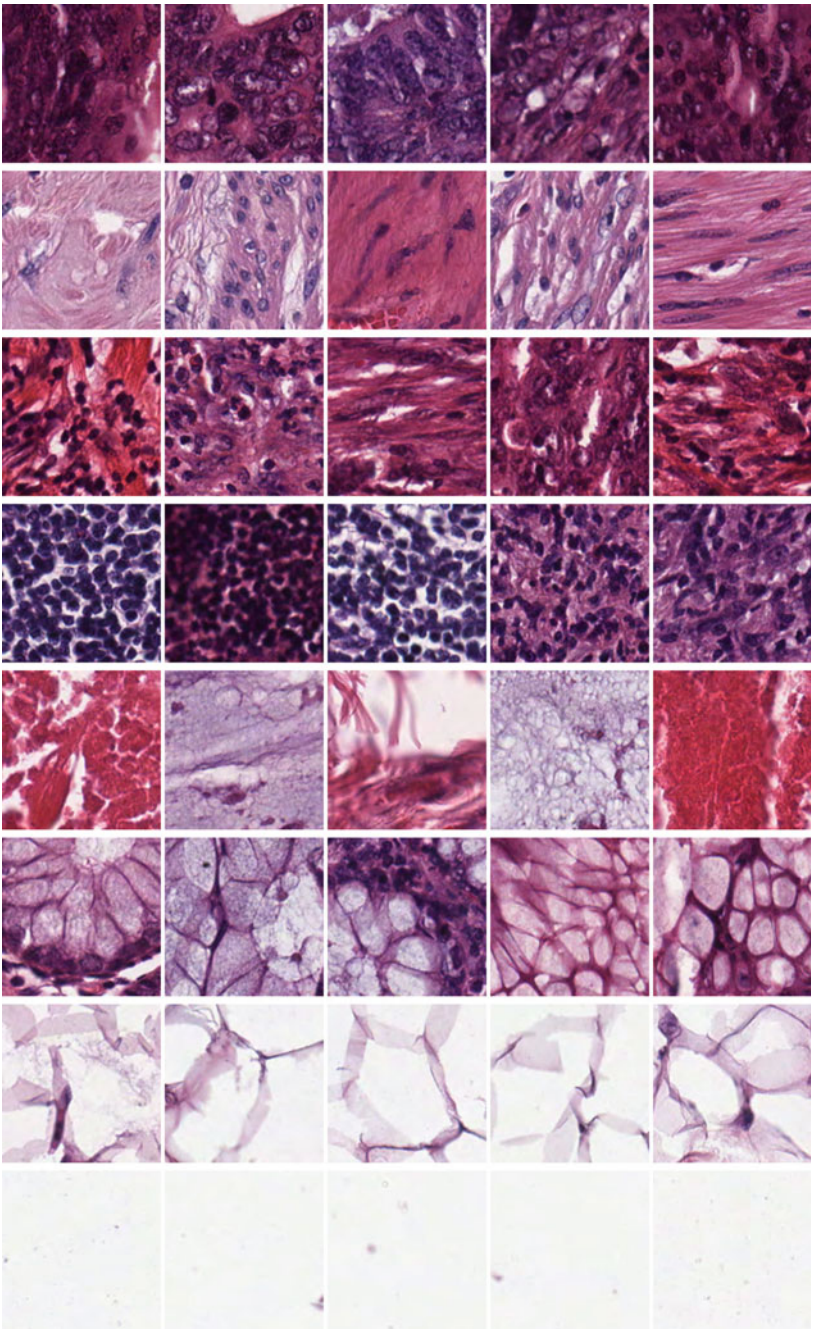


Fig. 3.4 Multi-class Kather's dataset. From top to bottom: epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosa glands, adipose tissue and background

tional License, and the data can be accessed through Ref. [57]. Sample images are available in Fig. 3.4.

3.3.2.2 Multi-class LUMC

This dataset contains 274 images representing the following four tissue classes: *epithelium* (50 images), *stroma* (194), *lymphocytes* (15) and *necrosis* (15). The images come from 16 digitised slides of H&E stained colon tissue samples acquired from stage II and III colon cancer patients who underwent surgery at Leiden University Medical Center (Leiden, The Netherlands) between 2002 and 2010. The slides were digitized using a Philips IntelliSite Ultra Fast Scanner. The approximate spatial resolution ranges between 3.1 and 5.4 μm/px. Further details about the patient series, histopathological protocol and ethical guidelines are available in [58]. As in Multi-class Kather's dataset, the original slides show rather dissimilar appearance reflecting the variability in the preparation and acquisition process. The digital slides were manually segmented by two operators into homogeneous regions representing well-defined areas of the above-mentioned classes. Square tiles of dimension 90 px

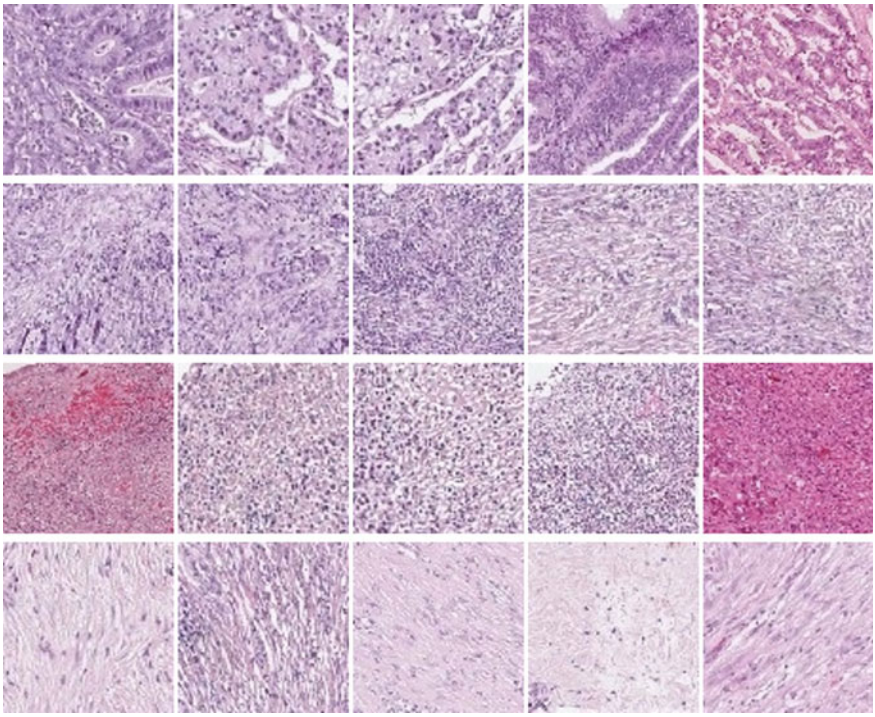


Fig. 3.5 Multi-class LUMC dataset. From top to bottom: epithelium, lymphocites, necrosis and stroma

× 90 px were finally cropped from the segmented images ensuring that at least 90% of the area of each tile was covered by the same class of tissue. Sample images are shown in Fig. 3.5.

3.4 Methods

We considered 10 feature vectors from pre-trained deep network and 12 LBP-variants as described below.

3.4.1 Features from Pre-trained Convolutional Networks

An array of feature vectors from the 10 pre-trained models listed below was obtained by taking the L_2-normalised output of the last fully-connected layer of each network (this is usually referred to as the 'FC' configuration [59]). Preprocessing was limited to resizing the input images to the networks' receptive field. The number of features was 4906 for the Caffe-Alex and VGG networks, 2048 for the ResNets and 1024 for the GoogLeNet. Apart from VGG-Face, which was trained on a *ad hoc* dataset of faces [60], all the other models were trained for object recognition tasks on the ImageNet dataset [61].

3.4.1.1 Pre-trained Models

Caffe-Alex A simple network architecture composed of five convolutional and three fully-connected layers [20].

GoogLeNet A deep network architecture including 22 learnable layers based on the 'Inception' model as described in [62].

Caffe-Alex A simple network architecture composed of five convolutional and three fully-connected layers [20].

GoogLeNet A deep network architecture including 22 learnable layers implementing the 'Inception' model as described in [62].

ResNet-50 and **ResNet-101** Two very deep networks trained by residual learning respectively containing 50 and 101 layers [63].

VGG-Face A convolutional network for face recognition tasks including eight convolutional and three fully-connected layers [60].

VGG-F, **VGG-M** and **VGG-S** Three networks with the same structure as Caffe-Alex's but differing in the size of the convolutional masks, stride and pooling window [64].

VGG-VeryDeep-16 and **VGG-VeryDeep-19** Two deep networks extending the VGGs models respectively featuring 13 convolutional plus three fully-connected layers, and 13 convolutional plus six fully-connected layers [65].

3.4.2 LBP Variants

Local image features based on LBP variants were computed using the following descriptors[1]:

- Binary Gradient Contours (BGC) [66]
- Completed Local Binary Patterns (CLBP) [67]
- Extended Local Binary Patterns (ELBP) [68]
- Gradient-Based Local Binary Patterns (GLBP) [69]
- Improved Local Binary Patterns (ILBP) [70]
- Improved Opponent Colour Local Binary Patterns (IOCLBP) [71]
- Local Binary Patterns (LBP) [72]
- Local Colour Vector Binary Patterns (LCVBP) [73]
- Local Ternary Patterns (LTP) [74]
- Opponent Colour Local Binary Patterns (OCLBP) [75]
- Rank Transform (RT) [76]
- Texture Spectrum (TS) [77].

For each of the above we computed both the directional (orientation-selective) and rotation-invariant features, respectively indicated as 'dir' and 'ri' in the remainder. Invariance against rotation was obtained by grouping together the local patterns that can be transformed into one another through a discrete rotation [72]. Note that RT features are rotation-invariant by definition. For each descriptor a multi-resolution feature vector was obtained by concatenating the vectors computed at resolution 1, 2, and 3 px (same settings as in [78]). The number of features ranged from a minimum of 27 (RT) to a maximum of 19683 (TS, 'dir').

3.5 Experiments

To evaluate the effectiveness of the image descriptors described in Sect. 3.4 we carried out a set of supervised classification experiments on the datasets detailed in Sect. 3.3. For classification we used two different strategies:

[1]Underline indicates colour descriptors. For a detailed description of each method please refer to the given references.

1. Non-parametric nearest-neighbour (1-NN) rule with cityblock L_1 distance
2. Support Vector Machines (SVM) with radial-basis kernel functions.

This choice was based on the consideration that, on the one hand, 1-NN always provides a good experimental baseline for its being parameter-free and conceptually simple: this is the reason why it is commonly used for comparative purposes [28, 59, 79]. On the other hand, tunable methods such as SVM can usually achieve substantial improvement on the 1-NN baseline, therefore give a better estimate of the real potential of the image descriptors tested. In this case SVM's penalty and regularisation parameter (respectively C and γ in the remainder) were estimated beforehand through a grid search procedure [80] over the following sets of values: $C \in \{2^{-2}, 2^0, \ldots, 2^8\}$ and $\gamma \in \{2^{-8}, 2^{-6}, \ldots, 2^8\}$. The search returned $C = 2^8$ and $\gamma = 2^8$ as optimal values for LBP variants, and $C = 2^8$ and $\gamma = 2^0$ for features from pre-trained convolutional networks.

Accuracy estimation was based on split-sample validation with stratified sampling and a training ratio of 50%—i.e.: for each dataset half of the labelled tissue patches of each class (half plus one when the number was odd) were used to train the classifier and the remaining half to test it. To stabilise the estimation the random subdivision into training and test set was repeated 100 times, which generated $P = 100$ classification problems for each dataset. For each classification problem p the accuracy a_p was the fraction of tissue patches correctly classified:

$$a_p = \frac{S_c}{S} \tag{3.1}$$

where S_c is the number of hits and S the total number of tissue patches to classify. The overall accuracy $\hat{a}$ was computed as the average over the 100 problems:

$$\hat{a} = \frac{1}{P} \sum_{p=1}^{P} a_p \tag{3.2}$$

The resulting values of $\hat{a}$ by dataset, image descriptor and classification method are reported in Tables 3.1 and 3.2. For each dataset the highest accuracy achieved by LBP variants and CNN-based feature are indicated in boldface. When there was, within the same dataset, a statistically significant difference between the best CNN-based feature vector and the best LBP variant the highest figure was highlighted in grey. Statistical significance was tested through non-parametric Wilcoxon-Mann-Whitney rank sum test [81] over the set of accuracy values resulting from the 100 subdivisions into train and test set.

Table 3.1 Average classification accuracy ($\hat{a}$, in %). Classifier: 1-NN

Descriptor	Datasets							Average
	1	2	3	4	5	6	7	
CNN-based features								
Caffe-Alex-FC	94.7	87.8	97.9	94.5	94.4	97.8	83.4	92.9
VGG-Face-FC	92.7	83.2	97.0	93.5	93.5	96.6	78.3	90.7
VGG-F-FC	96.5	89.1	98.2	94.1	94.9	98.4	85.7	93.8
VGG-M-FC	95.9	88.8	98.2	94.1	95.0	97.0	85.2	93.5
VGG-S-FC	96.8	90.1	98.7	94.7	94.4	98.0	86.2	94.1
VGG-VD-16-FC	97.4	91.5	98.5	95.1	95.1	97.4	86.1	94.4
VGG-VD-19-FC	95.4	88.9	98.8	94.7	94.5	97.2	84.2	93.4
GoogleNet-FC	97.0	87.0	97.4	93.2	93.9	96.1	79.3	92.0
ResNet-50-FC	**98.7**	**93.4**	**99.0**	96.2	**97.3**	98.7	**89.7**	96.2
ResNet-101-FC	98.3	91.4	98.8	**97.4**	96.6	**98.9**	89.3	95.8
LBP variants (dir/ri)								
BGC	96.1/96.5	85.2/83.9	95.1/93.2	88.4/89.1	90.1/87.8	90.7/88.6	71.9/71.5	88.2/87.2
CLBP	98.8/**99.1**	91.5/92.4	96.0/94.6	88.9/88.3	94.9/**97.2**	94.9/95.6	81.7/86.0	92.4/93.3
GLBP	98.8/98.5	92.1/90.2	95.3/94.7	89.4/86.6	91.2/91.7	93.7/92.0	78.3/78.3	91.3/90.3
ILBP	98.5/98.8	91.2/91.7	94.9/92.3	88.3/87.5	92.1/92.4	94.2/94.6	78.7/82.3	91.1/91.4
LBP	96.4/96.0	86.9/84.3	95.0/93.2	88.4/87.3	91.5/92.3	91.3/93.4	72.1/75.5	88.8/88.9
LTP	96.7/96.8	86.8/85.1	95.0/93.0	88.8/87.5	91.6/92.7	91.1/93.5	72.5/76.6	88.9/89.3
RT	96.1	82.1	91.4	87.2	92.5	91.3	72.5	87.6
TS	96.4/97.5	84.0/85.1	96.0/96.2	90.3/91.6	91.0/91.8	92.5/93.8	75.5/80.0	89.4/90.9
ELBP	96.4/96.5	87.8/84.8	95.3/94.4	88.9/89.4	91.6/92.4	91.6/92.8	74.4/78.2	89.4/89.8
OCLBP	98.8/**99.1**	94.1/92.9	97.8/96.8	92.2/91.5	92.7/91.9	98.3/98.1	89.5/91.0	94.8/94.5
IOCLBP	98.8/99.0	94.6/**94.8**	**97.9**/97.2	92.5/92.9	92.6/91.6	**98.7**/98.6	90.2/**92.4**	95.0/95.2
LCVBP	98.0/98.2	92.1/91.4	96.6/97.3	91.2/**94.1**	92.6/91.9	98.1/97.8	88.8/91.0	93.9/94.5

Datasets: (1) LUMC-TwoClass; (2) LUMC-MultiClass; (3) NKI; (4) VGH; (5) Epistroma; (6) Kather-TwoClass; (7) Kather-MultiClass. Figures in boldface indicate the best result for each dataset and class of method, grey cells a statistically significant difference ($p < 0.05$)

3.6 Results and Discussion

Table 3.1 and Fig. 3.6 report the overall accuracy of CNN-based features and LBP variants for the NN-L1 classifier, while Table 3.2 and Fig. 3.7 show the performance with the SVM classifier.

With the nearest-neighbour classifier the results were quite levelled: LBP variants outperformed CNN-based features in three datasets out of seven, the reverse occurred in another three datasets, whereas in the remaining one the difference did not reach statistical significance. With SVM the outcome was decidedly more favourable to CNN-based features, which outperformed LBP variants in six dataset out of seven,

Table 3.2 Average classification accuracy ($\hat{a}$, in %). Classifier: SVM

Descriptor	Datasets							Average
	1	2	3	4	5	6	7	
CNN-based features								
Caffe-Alex-FC	97.9	91.9	99.2	96.9	97.1	99.1	90.9	96.1
VGG-Face-FC	96.7	89.8	98.5	95.0	95.9	98.3	87.1	94.5
VGG-F-FC	98.4	92.0	99.1	96.0	98.0	99.0	92.2	96.4
VGG-M-FC	98.5	92.2	98.9	96.3	97.9	98.8	91.4	96.3
VGG-S-FC	97.7	92.8	99.1	95.5	97.6	98.9	91.7	96.2
VGG-VD-16-FC	98.3	93.9	99.3	95.9	98.4	98.5	91.6	96.6
VGG-VD-19-FC	97.3	92.4	99.3	95.7	97.9	98.6	90.9	96.0
GoogleNet-FC	98.6	90.2	98.7	94.3	97.3	98.7	88.4	95.2
ResNet-50-FC	**99.2**	**95.2**	**99.6**	98.0	**98.9**	**99.4**	**94.6**	97.8
ResNet-101-FC	99.1	94.1	99.5	**98.4**	98.6	**99.4**	94.1	97.6
LBP variants (dir/ri)								
BGC	96.9/97.0	88.9/87.1	96.9/95.4	91.6/93.9	94.5/95.1	96.3/93.9	83.7/78.5	92.7/91.5
CLBP	98.9/**99.3**	93.4/93.4	97.7/98.2	92.3/93.2	97.1/97.3	97.3/97.1	88.6/89.5	95.0/95.4
GLBP	**99.3**/98.3	93.2/90.6	97.1/96.2	91.0/90.9	94.0/94.9	96.8/94.4	86.5/83.9	94.0/92.7
ILBP	98.6/98.5	92.4/92.9	96.4/95.8	90.1/89.6	94.9/95.6	96.7/96.2	87.5/88.0	93.8/93.8
LBP	97.4/96.8	89.6/87.3	96.7/95.8	90.3/91.2	94.3/94.4	96.2/95.1	83.5/82.6	92.6/91.9
LTP	97.6/97.0	90.5/87.5	97.2/96.0	91.5/92.2	95.0/95.8	96.8/95.6	85.4/84.8	93.4/92.7
RT	96.7	83.8	93.5	88.9	94.4	92.6	77.3	89.6
TS	97.9/97.7	91.5/89.7	97.7/97.4	92.7/94.2	94.7/95.3	96.9/96.1	86.1/86.3	93.9/93.8
ELBP	96.9/96.9	89.3/88.4	96.6/96.6	89.7/92.2	93.9/95.1	95.1/95.7	84.2/85.2	92.3/92.9
OCLBP	98.8/99.1	**93.6/93.6**	98.0/97.8	93.4/94.7	92.3/93.7	98.6/98.6	92.6/93.5	95.3/95.9
IOCLBP	97.6/98.3	93.0/93.0	97.6/97.4	92.8/95.2	92.2/93.4	98.5/98.5	92.2/**93.7**	94.8/95.7
LCVBP	97.9/98.2	92.7/91.9	**98.6**/98.1	**95.7**/95.6	97.0/**97.4**	**99.2**/98.9	93.5/93.4	96.4/96.2

Datasets: (1) LUMC-TwoClass; (2) LUMC-MultiClass; (3) NKI; (4) VGH; (5) Epistroma; (6) Kather-TwoClass; (7) Kather-MultiClass. Figures in boldface indicate the best result for each dataset and class of method, grey cells a statistically significant difference ($p < 0.05$)

whereas in the remaining one there was no significant difference. The absolute accuracy values were on average better with SVM—as one would expect.

ResNet-50 and ResNet-101 proved the best pre-trained models with all the datasets and regardless of the classifier used, whereas IOCLBP, OCLBP and LCVBP emerged as the best LBP variants. However, no clear trend emerged as concerns the relative performance of directional versus rotation-invariant features.

To put our results in the context of the current literature, Table 3.3 shows the accuracy values published in recent related works. It is to be noted, however, that the table lends itself to a qualitative comparison only, for the classifiers and accuracy estimation settings used in the cited references differ from one study to another. In any case the figures confirm the effectiveness of the image descriptors studied here, particularly CNN-based features. These results parallel those in [25, 82], suggesting

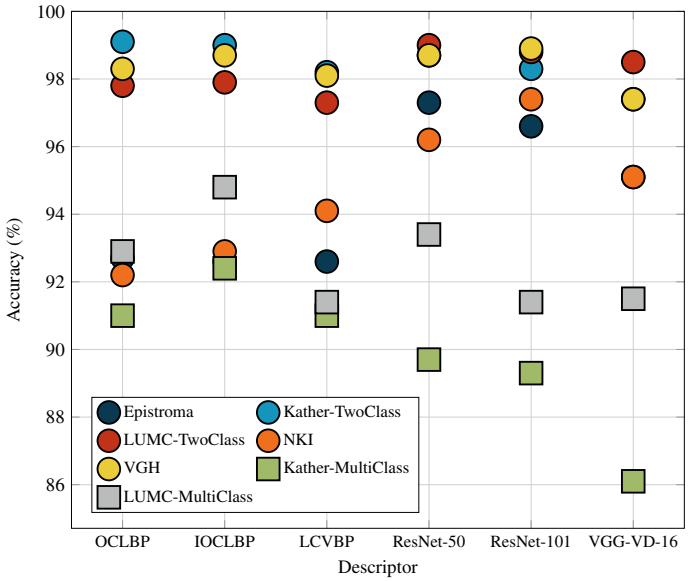


Fig. 3.6 Overall accuracy of the best image descriptors. Classifier: 1-NN (L_1). Circular markers indicate two-class datasets, square markers multi-class datasets

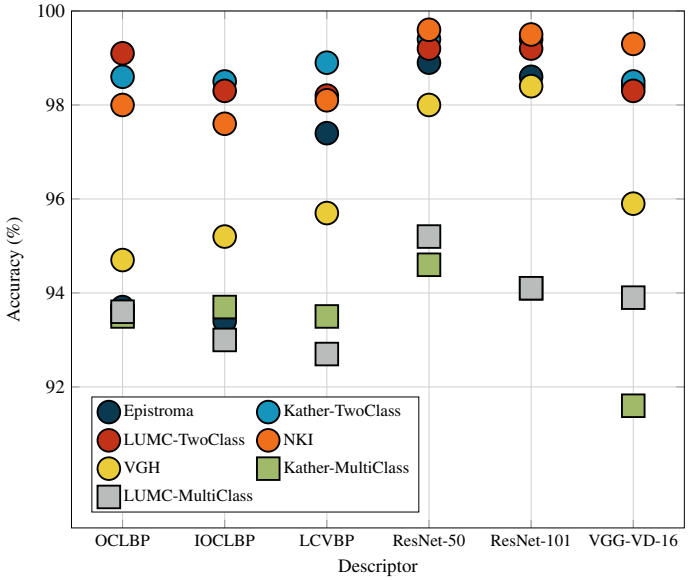


Fig. 3.7 Overall accuracy of the best image descriptors. Classifier: SVM. Circular markers indicate two-class datasets, square markers multi-class datasets

Table 3.3 Comparison with related works. Accuracy values are in %

Reference	Yea	Dataset	Accuracy reported	Best accuracy (this work)	Best method (this work)
Linder et al. [17]	2012	Epistroma	96.8	**98.9**	ResNet-50-FC
Bianconi et al. [37]	2015	Epistroma	96.9	**98.9**	ResNet-50-FC
Nava et al. [38]	2016	Epistroma	96.5	**98.9**	ResNet-50-FC
Xu et al. [44]	2016	Epistroma	**100**	98.9	ResNet-50-FC
		NKI	84.3	**99.6**	ResNet-50-FC
		VGH	88.3	**98.4**	ResNet-101-FC
Kather et al. [39]	2016	Two-class Kather's	98.6	**99.4**	ResNet-50-FC
		Multi-class Kather's	87.4	**94.6**	ResNet-50-FC
Huang et al. [83]	2017	NKI	90.6	**99.8**	ResNet-50-FC
		VGH	94.4	**98.4**	ResNet-101-FC

that image features from pre-trained CNN can in most cases outperform traditional, hand-designed descriptors like LBP variants. The capability of pre-trained convolutional networks to be used in context different from those they have been trained on (transfer learning) was at the beginning a surprise to many [26], but this work once again confirms the effectiveness of this approach. We believe this difference is mainly related to the ability of CNN-based features—particularly the 'FC' configuration—to capture not only the statistic distribution of local image patterns, but also their spatial interaction—which is otherwise neglected by LBP variants.

3.7 Conclusions

The visual assessment of histopathological images including the classification of tissue regions into meaningful classes plays a pivotal role in the management of patients with neoplastic disorders. Nowadays digital slide scanners have made this once typically manual procedure amenable to being addressed by automated image processing approaches. In this chapter we have investigated the potential offered by two classes of image descriptors: pre-trained convolutional neural networks and LBP variants. Both offer interesting advantages in that they are computationally cheap, easy to use and can be seamlessly integrated with standard classifiers.

Experimenting on seven different datasets of histopathological images we demonstrated that both classes of methods can be effective for the task, though a noticeable superiority emerged in favour of features from convolutional networks. In particular, we found that these can be seamlessly integrated with standard classifiers (e.g. Support Vector Machines) to attain overall discrimination between 95 and 99%. Notably, the approach proved fairly stable throughout all the datasets used—despite

the images being different in a number of features such as spatial resolution, size and image acquisition procedure.

In light of the complexities in solid tumours, variability between the cellular content in different cancer entities can be addressed with this approach. As the datasets show, a considerable variation in stromal content is robustly detected and reliably measured across diverse sets of histology images. With regard to image acquisition and tissue preparation, no direct limitation was detectable. Further studies will elucidate the clinical applicability of this approach and the possible usability within a medical context.

Acknowledgements This work was partially supported by the Italian Ministry of University and Research (MIUR) under the Individual Funding Scheme for Fundamental Research 'FFABR' 2017 (F. Bianconi) and by the Department of Engineering at the Università degli Studi di Perugia, Italy, within the Fundamental Research Grants Scheme 2018 (F. Bianconi).

References

1. Huijbers, A., Tollenaar, R.A.E.M., Pelt, G.W.V., Zeestraten, E.C.M., Dutton, S., McConkey, C.C., Domingo, E., Smit, V.T.H.B.M. , Midgley, R., Warren, B.F., Johnstone, E.C., Kerr, D.J., Mesker, W.E.: The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. Ann. Oncol. **24**(1), 179–185 (2013)
2. Park, J.H., Richards, C.H., McMillan, D.C., Horgan, P.G., Roxburgh, C.S.D.: The relationship between tumour stroma percentage, the tumour microenvironment and survival in patients with primary operable colorectal cancer. Ann. Oncol. **25**(3), 644–651 (2014)
3. van Pelt, G.W., Sandberg, T.P., Morreau, H., Gelderblom, H., van Krieken, J.H.J.M., Tollenaar, R.A.E.M., Mesker, W.E.: The tumour-stroma ratio in colon cancer: the biological role and its prognostic impact. Histopathology **73**(2), 197–206 (2018). August
4. Chen, Y., Zhang, L., Liu, W., Liu, X.: Prognostic significance of the tumor-stroma ratio in epithelial ovarian cancer. BioMed Res. Int. **2015** (2015)
5. Lv, Z., Cai, X., Weng, X., Xiao, H., Du, C., Cheng, J., Zhou, L., Xie, H., Sun, K., Wu, J., Zheng, S.: Tumor-stroma ratio is a prognostic factor for survival in hepatocellular carcinoma patients after liver resection or transplantation. Surgery **158**(1), 142–150 (2015)
6. Zhang, X.-L., Jiang, C., Zhang, Z.-X., Liu, F., Zhang, F., Cheng, Y.-F.: The tumor-stroma ratio is an independent predictor for survival in nasopharyngeal cancer. Oncol. Res. Treat. **37**(9), 480–484 (2014)
7. De Kruijf, E.M., van Nes, J.G.H., van De Velde,C.J.H., Putter, H., Smit, V.T.H.B.M., Liefers, G.J., Kuppen, P.J.K., Tollenaar, R.A.E.M., Mesker, W.E.: Tumor-stroma ratio in the primary tumor is a prognostic factor in early breast cancer patients, especially in triple-negative carcinoma patients. Breast Cancer Res. Treat. **125**(3), 687–696 (2011)
8. Dekker, T.J.A., van De Velde, C.J.H., van Pelt, G.W., Kroep, J.R., Julien, J.-P., Smit, V.T.H.B.M., Tollenaar, R.A.E.M., Mesker, W.E.: Prognostic significance of the tumor-stroma ratio: validation study in node-negative premenopausal breast cancer patients from the eortc perioperative chemotherapy (pop) trial (10854). BBreast Cancer Res. Treat. **139**(2):371–379 (2013)
9. Vangangelt, K.M.H., van Pelt, G.W., Engels, C.C., Putter, H., Liefers, G.J., Smit, V.T.H.B.M., Tollenaar, R.A.E.M., Kuppen, P.J.K., Mesker, W.E.: Prognostic value of tumor–stroma ratio

combined with the immune status of tumors in invasive breast carcinoma. Breast Cancer Res. Treat. **168**(3), 601–612 (2018)
10. Mouawad, R., Spano, J.-P., Khayat, D.: Lymphocyte infiltration in breast cancer: a key prognostic factor that should not be ignored. J. Clin. Oncol. **29**(33), 4471 (2011)
11. Correale, P., Rotundo, M.S., Botta, C., Vecchio, M.T.D., Tassone, P., Tagliaferri, P.: Tumor infiltration by chemokine receptor 7 (ccr7)+ t-lymphocytes is a favorable prognostic factor in metastatic colorectal cancer. OncoImmunology **1**(4), 531–532 (2012)
12. Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P.H., Trajanoski, Z., Fridman, W.H., Pagès, F.: Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science **313**(5795), 1960–1964 (2006). September
13. Halama, N., Michel, S., Kloor, M., Zoernig, I., Benner, A., Spille, A., Pommerencke, T., von Knebel, D.M., Folprecht, G., Luber, B., Feyen, N., Martens, U.M., Beckhove, P., Gnjatic, S., Schirmacher, P., Herpel, E., Weitz, J., Grabe, N., Jaeger, D.: Localization and density of immune cells in the invasive margin of human colorectal cancer liver metastases are prognostic for response to chemotherapy. Cancer Res. **71**(17), 5670–5677 (2011)
14. Ness, N., Andersen, S., Valkov, A., Nordby, Y., Donnem, T., Al-Saad, S., Busund, L.-T., Bremnes, R.M., Richardsen, E.: Infiltration of cd8+ lymphocytes is an independent prognostic factor of biochemical failure-free survival in prostate cancer. Prostate **74**(14), 1452–1461 (2014)
15. Caruso, R., Parisi, A., Bonanno, A., Paparo, D., Emilia, Q., Branca, G., Scardigno, M., Fedele, F.: Histologic coagulative tumour necrosis as a prognostic indicator of aggressiveness in renal, lung, thyroid and colorectal carcinomas: a brief review. Oncol. Lett. **3**(1), 16–18 (2012)
16. Hynes, S.O., Coleman, H.G., Kelly, P.J., Irwin, S., O'Neill, R.F., Gray, R.T., Mcgready, C., Dunne, P.D., Mcquaid, S., James, J.A., Salto-Tellez, M., Loughrey, M.B.: Back to the future: routine morphological assessment of the tumour microenvironment is prognostic in stage ii/iii colon cancer in a large population-based study. Histopathology **71**(1), 12–26 (2017). In press
17. Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J.: Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagn. Pathol. **7**(22), 1–11 (2012)
18. Courrech Staal, E.F.W., Smit, V.T.H.B.M., van Velthuysen, M.-L.F., Spitzer-Naaykens, J.M.J., Wouters, M.W.J.M., Mesker,W.E., Tollenaar, R.A.E.M., van Sandick, J.W.: Reproducibility and validation of tumour stroma ratio scoring on oesophageal adenocarcinoma biopsies. Eur. J. Cancer **47**(3), 375–382 (2011)
19. Fuchs, T.J., Buhmann, J.M.: Computational pathology: challenges and promises for tissue analysis. Comput. Med. Imaging Graph. **35**(7–8), 515–530 (2011)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, USA (December 2012)
21. Sang, H., Zhou, Z.: Automatic detection of human faces in color images via convolutional neural networks. ICIC Express Lett., Part B: Appl. **7**(4) (2016)
22. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Fisher networks for large-scale image classification. In:Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, USA (December 2013)
23. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
24. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017). February
25. Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A.: Deep filter banks for texture recognition, description, and segmentation. Int. J. Comput. Vis. **118**(1), 65–94 (2016)
26. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2014, pp. 512–519, Columbus, USA (June 2014)

27. Fernández, A., Álvarez, M.X., Bianconi, F.: Texture description through histograms of equivalent patterns. J. Math. Imaging Vis. **45**(1), 76–102 (2013)
28. Liu, L., Fieguth, P., Guo, Y., Wang, X., Pietikäinen, M.: Local binary features for texture classification: taxonomy and experimental study. Pattern Recognit. **62**, 135–160 (2017)
29. Meijer, G.A., Beliën, J.A.M., Van Diest, P.J., Baak, J.P.A.: Image analysis in clinical pathology. J. Clin. Pathol. **50**(5), 365–370 (1997)
30. Al-Janabi, S., Huisman, A., Van Diest, P.J.: Digital pathology: current status and future perspectives. Histopathology **61**(1), 1–9 (2012)
31. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
32. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inform. **7**(1) (2016). Art. no. 29
33. Madabhushi, A., Lee, G.: Image analysis and machine learning in digital pathology: challenges and opportunities. Med. Image Anal. **33**, 170–175 (2016)
34. Veta, M., Pluim, J.P.W., Van Diest, P.J., Viergever, M.A.: Breast cancer histopathology image analysis: a review. IEEE Trans. Biomed. Eng. **61**(5), 1400–1411 (2014)
35. Watanabe, K., Kobayashi, T., Wada,T.: Semi-supervised feature transformation for tissue image classification. PLoS ONE **11**(12) (2016). Article number e0166413
36. Diamond, J., Anderson, N.H., Bartels, P.H., Montironi, R., Hamilton, P.W.: The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. Hum. Pathol. **35**(9), 1121–1131 (2004)
37. Bianconi, F., Fernández, A., Álvarez Larrán, A.: Discrimination between tumour epithelium and stroma via perception-based features. Neurocomputing **154**, 119–126 (2015)
38. Nava, R., González, G., Kybic, J., Escalante-Ramírez, B.: Classification of tumor epithelium and stroma in colorectal cancer based on discrete Tchebichef moments. Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9401**, 79–87 (2016)
39. Kather, J.N., Weis, C.-A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G.: Multi-class texture analysis in colorectal cancer histology. Sci. Rep. **6** (2016). Art. no. 27988
40. Badejo, J.A., Adetiba, E., Akinrinmade, A., Akanle, M.B.: Medical image classification with hand-designed or machine-designed texture descriptors: a performance evaluation. In: Rojas, I., Ortuño, F. (eds.) Proceedings of the International Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Lecture Notes in Computer Science, vol. 10814, Granada, Spain, April 2018, pp. 266–275. Springer
41. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging **35**(5), 1153–1159 (2016)
42. Ciompi, F., Geessink, O., Bejnordi, B.E., de Souza, G.S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I., van der Laak, J.: The importance of stain normalization in colorectal tissue classification with convolutional networks. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, Melbourne, Australia, April 2017 (2017). To appear
43. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: Breast cancer histopathological image classification using convolutional neural networks. In: Proceedings of the International Joint Conference on Neural Networks, pp. 2560–2567, Vancouver, Canada (July 2016)
44. Xu, J., Luo, X., Wang, G., Gilmore, H., Madabhushi, A.: A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing **191**, 214–223 (2016)
45. Gao, Z., Wang, L., Zhou, L., Zhang, J.: HEp-2 cell image classification with deep convolutional neural networks. IEEE J. Biomed. Health Inform. **21**(2), 416–428 (2017). March
46. Veta, M., van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B.L., Vestergaard, J.S., Dahl, A.B., Cireşan, D.C., Schmidhuber, J., Giusti, A., Gambardella, L.M., Tek, F.B., Walter, T., Wang, C.-W., Kondo, S., Matuszewski, B.J., Precioso, F., Snell, V., Kittler, J., de Campos, T.E., Khan, A.M., Rajpoot, N.M., Arkoumani, E., Viergever,

M.A, Lacle, M.M., Pluim, J.P.W.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical Image Analysis **20**(1):237–248 (2015)
47. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
48. Van Ginneken, B., Setio, A.A.A., Jacobs, C., Ciompi,F.: Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: Proceedings of the 12th IEEE International Symposium on Biomedical Imaging, ISBI 2015, pp. 286–289, Brooklyn, USA (April 2015)
49. Arevalo, J., Gonzalez, F.A., Ramos-Pollan, R., Oliveira, J.L., Lopez, M.A.G.: Convolutional neural networks for mammography mass lesion classification. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Milan, Italy, November 2015, pp. 797–800
50. Vedaldi, A., Lenc, K.: MatConvNet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM International Conference on Multimedia (MM 2015), pp. 689–692, Brisbane, Australia (October 2015)
51. Jia, J., Shelhamer, E.: Caffe deep learning framework. http://caffe.berkeleyvision.org/. Last accessed 19 Apr 2017
52. Webmicroscope: EGFR colon TMA stroma LBP classification (2012). http://fimm.webmicroscope.net/Research/Supplements/epistroma. Accessed 17 Aug 2018
53. Linder, N., Martelin, E., Lundin, M., Louhimo, J., Nordling, S., Haglund, C., Lundin, J.: Xanthine oxidoreductase—clinical significance in colorectal cancer and in vitro expression of the protein in human colon cancer cells. Eur. J. Cancer **45**(4), 648–655 (2009). March
54. Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M.J., West, R.B., van de Rijn, M., Koller, D.: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. **3**(108), 1–11 (2011)
55. Beck, A., Sangoi, A., Leung, S., Marinelli, R., Nielsen, T., van de Vijver, M., West, R., van de Rijn, M., Koller, D.: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival (2011). https://tma.im/tma-portal/C-Path/images.html. Accessed 2 Mar 2017
56. Kather, J.N., Marx, A., Reyes-Aldasoro, C.C., Schad, L.R., Zöllner, F.G., Weis, C.A.: Continuous representation of tumor microvessel density and detection of angiogenic hotspots in histological whole-slide images. Oncotarget **6**(22), 19163–19176 (2015). August
57. Kather, J. N., Zöllner, F. G., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., Weis, C.-A.: Collection of textures in colorectal cancer histology (May 2016)
58. Mesker, W.E., Junggeburt, J.M.C., Szuhai, K., De Heer, P., Morreau, H., Tanke, H.J., Tollenaar, R.A.E.M.: The carcinoma-stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage. Cell. Oncol. **29**(5), 387–398 (2007)
59. Cusano, C., Napoletano, P., Schettini, R.: Combining multiple features for color texture classification. J. Electron. Imaging **25**(6), (2016). Article number 061410
60. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015, Swansea, UK (2015)
61. ImageNet. http://www.image-net.org. Accessed 23 Feb 2018
62. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR2015), Boston, USA (2015)
63. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016)
64. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference 2014, Nottingham, United Kingdom (2014)
65. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In:Proceedings of the 5th International Conference on Learning Representations, San Diego, USA (2015)

66. Fernández, A., Ghita, O., González, E., Bianconi, F., Whelan, P.F.: Evaluation of robustness against rotation of lbp, ccr and ilbp features in granite texture classification. Mach. Vis. Appl. **22**(6), 913–926 (2011)
67. Guo, Z., Zhang, L.: A completed modeling of local binary patterns operator for texture classification. IEEE Trans. Image Process. **19**(6), 1657–1663 (2010)
68. Liu, L., Zhao, L., Long, Y., Kuang, G., Fieguth, P.: Extended local binary patterns for texture classification. Image Vis. Comput. **30**, 86–99 (2012)
69. He, Y., Sang, N., Robust illumination invariant texture classification using gradient local binary patterns. In Proceedings of, : International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping, p. 2011. Xiamen, China (2011)
70. Jin, H., Liu, Q., Lu, H., Tong, X.: Face detection using improved LBP under Bayesian framework. In: Proceedings of the 3rd International Conference on Image and Graphics, pp. 306–309, Hong Kong, China (December 2004)
71. Bianconi, F., Bello-Cerezo, R., Napoletano, P.: Improved opponent color local binary patterns: an effective local image descriptor for color texture classification. J. Electron. Imaging **27**(1) (2018). Art. No. 011002
72. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
73. Lee, S., Choi, J., Ro, Y., Plataniotis, K.: Local color vector binary patterns from multichannel face images for face recognition. IEEE Trans. Image Process. **21**(4), 2347–2353 (2012)
74. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: Analysis and Modelling of Faces and Gestures, vol. 4778. Lecture Notes in Computer Science. Springer (2007)
75. Mäenpää, T., Pietikäinen, M.: Texture analysis with local binary patterns. In: Chen, C.H., Wang, P.S.P. (eds) Handbook of Pattern Recognition and Computer Vision, 3rd edn, pp. 197–216. World Scientific Publishing (2005)
76. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Proceedings of the 3rd European Conference on Computer Vision (ECCV 1994) (1994)
77. He, D.-C., Wang, L.: Texture unit, texture spectrum, and texture analysis. IEEE Trans. Geosci. Remote. Sens. **28**(4), 509–512 (1990)
78. Bianconi, F., Bello-Cerezo, R., Napoletano, P., Di Maria, F.: Improved opponent colour local binary patterns for colour texture classification. In: Bianco, S., Schettini, R., Tominaga, S., Tremeau, A. (eds.) Proceedings of the 6th Computational Color Imaging Workshop (CCIW 2017). Lecture Notes in Computer Science, vol. 10213, Milan, Italy, March 2017, pp. 272–281. Springer
79. Cernadas, E., Fernández-Delgado, M., González-Rufino, E., Carrión, P.: Influence of normalization and color space to color texture classification. Pattern Recognit. **61**, 120–138 (2017)
80. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification, 2016. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Last accessed 22 Mar 2017
81. Kanji, G.K.: 100 Statistical Tests, 3rd edn. Society for Industrial and Applied Mathematics (2006)
82. Cusano, C., Napoletano, P., Schettini, R.: Evaluating color texture descriptors under large variations of controlled lighting conditions. J. Opt. Soc. Am. A: Opt. Image Sci. Vis. **33**(1), 17–30 (2016)
83. Huang, Y., Zheng, H., Liui, C., Ding, X., Rohde, G.K.: Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE J. Biomed. Health Inform. **21**(6), 1625–1632 (2017)

Chapter 4
Ensemble of Handcrafted and Deep Learned Features for Cervical Cell Classification

Loris Nanni, Stefano Ghidoni, Sheryl Brahnam, Shaoxiong Liu and Ling Zhang

Abstract The aim of this work is to propose an ensemble of descriptors for Cervical Cell Classification. The system proposed here achieves strong discriminative power that generalizes well thanks to the combination of multiple descriptors based on different approaches, both learned and handcrafted. For each descriptor, a separate classifier is trained, then the set of classifiers is combined by sum rule. The system we propose here also presents a simple and effective method for boosting the performance of trained CNNs by combining the scores (using sum rule) of multiple CNNs into an ensemble. Different types of ensembles and different CNN topologies with different learning parameter sets are evaluated. Moreover, features extracted from tuned CNNs are used for training a set of Support Vector Machines (SVM). First, we validate our method on two cervical cell-related datasets; then, for more in-depth validation, we test the same system on other bioimage classification problems. Results show that the proposed system obtains state-of-the-art performance in all datasets, despite not being tuned on a specific dataset, i.e. the same descriptors with the same parameters are used in all the datasets. The MATLAB code of the descriptors will be available at https://github.com/LorisNanni.

L. Nanni (✉) · S. Ghidoni
Department of Information Engineering, University of Padua, via Gradenigo 6/B, 35131 Padua, Italy
e-mail: loris.nanni@unipd.it

S. Brahnam
Department of Information Technology and Cybersecurity, Glass Hall, Room 387, Missouri State University, 901 S. National, Springfield, MO 65804, USA
e-mail: sbrahnam@missouristate.edu

S. Liu
Department of Pathology, People's Hospital of Nanshan District, Shenzhen 518052, China

L. Zhang
Radiology and Imaging Sciences Department, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA
e-mail: zhangling0722@163.com

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_4

Keywords Deep learning · Ensemble of classifiers · Bioimage classification · Cancer data analysis

4.1 Introduction

Currently, image analysis is mainly performed by expert clinicians who provide the definitive assessments of medical images. The increasing number of medical images, however, demands automatic or semi-automatic analysis—an approach that is becoming more and more viable thanks to recent developments in the fields of image processing, pattern recognition, and image classification [1–3].

A significant part of bioimage processing relies on feature-based approaches, which are the best option when the local textured patterns provided by biological tissues are analyzed. A wide variety of texture features have thus been used in biomedical imaging classification systems: in some cases, ensembles of features are generated [4, 5] to increase the amount of extracted information. Some successful features include Gabor filters and Haralick textural features [6], which were developed in the early age of feature classification research, together with some more recent (and more widely used) descriptors, such as the scale-invariant feature transform (SIFT) [7] and Local Binary Patterns (LBP) [8], with their many variations [9]. Each of these features sets belong to what is called the *handcrafted* category since they have been designed by researchers to extract specific characteristics that are well described in the literature.

In contrast to handcrafted features are *learned* features, which, as the name suggests, are directly learned from data by a classifier system. As a result, these features tend to be very specific to the dataset used for training. However, if the training datasets are broad enough in range and contain a plethora of different images, the system is forced to learn a variety of patterns, making the learned features more generalizable as well. When this happens, learned features can be used alone, in the same way as handcrafted features are used. Learned features are increasingly being used for bioimage processing [10, 11]. Some examples of using learned features as standalone features can be found in [10] for the detection of ovarian carcinomas and in [11], where handcrafted and learned features are jointly employed.

In the last few years, a powerful learned approach has been proposed in the computer vision literature that is based on the deep learning paradigm [12]. Recent results indicate that deep learning is extremely effective in many image classification applications, with medical image analysis being counted among these successes [13]. Several papers have proposed methods to apply deep learning to a variety of medical image analysis tasks, including detection and counting (e.g. of mitotic events), segmentation (e.g. of nuclei), and tissue classification (e.g. of cancerous vs. non-cancerous samples) [14].

The most studied deep learning architecture is the Convolutional Neural Network (CNN) [15], a multi-layered image classification technique which incorporates spatial context and weight sharing between pixels and is able to learn the optimal image

features for a given classification task. CNN is inspired by the natural visual perception mechanism of human beings; it adopts effective representations of the original image and requires a small amount of preprocessing. The fundamental components of a CNN are several types of layers: convolutional, pooling, and fully-connected layers, whose weights are trained with the backpropagation algorithm. The deepest layers of the network act as feature extractors by evaluating sets of features in the input images that are learned directly from observations of the training set. The training phase, however, requires large volumes of labeled data to avoid the problem of over-fitting. When such training data are available [16], CNNs are capable of learning accurate and generalizable models; they are also capable of achieving state-of-the-art performance in general pattern recognition tasks.

A variety of CNN architectures have been introduced in the literature: among them are six representative architectures: LeNet [17], the first CNN proposed to classify handwritten digits; AlexNet [18], a deeper network designed for image classification; ZFNet [19], a newer model architecture that has been shown to outperform AlexNet; VGGNet [20], where some networks of increasing depth are achieved by using very small (3×3) convolution filters; GoogleNet [21], a deep CNN which includes "Inception" layers with improved utilization of the computing resources inside the network, and ResNet [22], a residual network which is easier to optimize than VGGNets.

When deep neural networks are trained on images, the first few convolution layers resemble either Gabor filters or color blobs that tend to be generalizable or transferable to many other image tasks and datasets [23]. Therefore, it is reasonable to consider pre-trained models as feature extractors that can then be used in the same way as handcrafted methods can be used, but with the difference being that the features extracted by a CNN are learned using the image dataset, whereas handcrafted descriptors are designed a priori by human experts to extract specific image characteristics. In contrast to the first convolution layers, features computed by the last layer of a pre-trained CNN network strictly depend on the dataset used for training and thus are specific to the classification problem. For this reason, a fine-tuning, or tuning, is required when applying the information contained in this layer to other datasets and classification problems.

In the literature there are three main methods using deep learning [24] to perform medical and bioimage classification: (1) training a CNN from scratch [25] using data preprocessing and selection to solve any imbalance in the data or any insufficient problems with it; (2) using transfer learning from a pre-trained CNN as a complementary feature extractor combined with existing handcrafted image features [26, 27]; and (3) using a pre-trained CNN and performing a tuning on target images [28, 29]. A fourth class of approaches can be defined considering the fusion of different CNN architectures: based on the observation that shallow CNNs may be too general and unable, therefore, to capture the subtle differences between highly variable images, while deep CNNs may be highly sensitive to subtle differences and unable to capture the general similarity between images. In [29] the authors propose an ensemble of fine-tuned CNNs pre-trained on a large dataset.

In this chapter, we present an ensemble of heterogeneous descriptors for bioimage classification expanding the tests detailed in [30]. The system proposed here combines descriptors based on both learned and handcrafted approaches. For each descriptor, a different classifier is trained. The set of classifiers and the classification results from the deep networks are then combined by sum rule.

Since one of our goals in this chapter is to investigate methods for building ensembles of CNNs by leveraging the classification power of pre-trained CNNs, we evaluate several different CNNs and assess them by running experiments using different learning rates, batch sizes, and topologies. We show that this simple ensemble approach produces a high performing, competitive system, one that outperforms the single best CNN trained specifically on the tested datasets. There are some disadvantages combining different CNNs, however. Although ensembles of CNNs perform exceptionally well, training such models requires considerable computational power (we used three TitanX GPUs in our experiments); and, since the total size of the network set is extremely large, substantially more computational power is required for input classification. All this computational power is needed in addition to that required to build an ensemble of SVMs by representing each image using features extracted from the different CNN layers.

Despite the initial computational costs, the main benefit of the proposed system is that it works well out-of-the-box, requiring little to no parameter tuning, yet performs competitively against less flexible systems optimized for particular image classification tasks and datasets. The reported results show that the proposed system obtains state-of-the-art performance in almost all the tested datasets, which represent a wide range of bioimage problems; our system achieves these results using the same set of descriptors across all the datasets.

4.2 Methods

As described in the introductory section, our methods include as well as evaluate both handcrafted and learned descriptors for cervical cell analysis and classification. This section briefly describes both types of descriptors. The following discussion divides descriptors into three categories: (1) handcrafted features (2) deep learned features produced by transfer learning, and (3) deep learned features produced by fine-tuning (FT).

Table 4.1 summarizes all the handcrafted descriptors used in our ensembles and provides the parameter sets used to extract each descriptor. Each is briefly described below this table. When SVM is used as classifier only the training data is used to fix its parameters.

All the tested datasets contain color images. For each handcrafted descriptor, which extracts features from a gray level input image, we run the feature extraction methods three times, once for each R, G, and B band. Descriptors extracted from each band are then used to train three SVMs that are finally combined by sum rule.

Table 4.1 Summary of handcrafted descriptors

Name	Parameters	Source
LTP	Multiscale Uniform LTP with two (R, P) configurations: (1, 8) and (2, 16), threshold = 3	[31]
MLPQ	Ensemble of LPQ descriptors obtained by varying the filter sizes, the scalar frequency, and the correlation coefficient between adjacent pixel values	[32]
CLBP	Completed LBP with two (R, P) configurations: (1, 8) and (2, 16)	[33]
RIC	Multiscale Rotation Invariant Co-occurrence of Adjacent LBP with $R \in \{1, 2, 4\}$	[34]
GOLD	Gaussian of Local Descriptors extracted using the spatial pyramid decomposition	[35]
HOG	Histogram of Oriented Gradients with 30 cells (5 by 6)	[36]
AHP	Adaptive Hybrid Pattern with *quantization level* = 5 and 2 (R, P) configurations: (1, 8) and (2, 16)	[37]
FBSIF	Extension of the BIF by varying the parameters of filter size (SIZE_BSIF, $size \in \{3, 5, 7, 9, 11\}$) and the threshold for binarizing (FULL_BSIF, $th \in \{-9, -6, -3, 0, 3, 6, 9\}$)	[38]
COL	A simple and compact color descriptor	[39]
MOR	A set of morphological features	[40]
CLM	Codebookless Model, we use the ensemble named CLoVo_3 in [17] based on e-SFT, PCA for dimensionality reduction and one-vs-all SVM for the training phase	[41]
LET	Same parameters fixed the source code of [42]	[42]

4.2.1 Handcrafted Features

4.2.1.1 LTP

LTP (Local Ternary Patterns) [31] (threshold = 3) is a texture descriptor belonging to the feature family derived from LBP (Local Binary Patterns) [8] that is designed to reduce the noise in the feature vector when uniform regions are analyzed.

4.2.1.2 MLPQ

MLPQ (Multithreshold LPQ) [32] is another feature belonging to the same LBP family: it extends the multi-threshold approach described for LBP to the LPQ feature that is based on the phase of the Short-Term Fourier Transform (STFT) evaluated on a rectangular neighborhood of size R. The MLPQ features used in our experiments are computed using parameter belonging to the following ranges: $\tau \in \{0.2, 0.4, 0.6, 0.8, 1\}$, $R \in \{1, 3, 5\}$, $a \in \{0.8, 1, 1.2, 1.4, 1.6\}$ and $\rho \in \{0.75, 0.95, 1.15, 1.35, 1.55, 1.75, 1.95\}$. Such sets were proposed in [43].

4.2.1.3 CLBP

Completed LBP (CLBP) [33] encodes texture by means of two components: the difference sign and the different magnitude computed between a reference pixel and all the pixels belonging to the neighborhood.

4.2.1.4 RIC

The multiscale Rotation Invariant Co-occurrence of Adjacent LBP (RIC) [34] considers the co-occurrence in the context of LBP features, i.e. the spatial relations among pixels. This feature adds rotational invariance for angles that are multiple of 45°. RIC depends on two parameters: LBP radius and displacement among the LBPs. The values used in our experiments are: (1, 2), (2, 4) and (4, 8).

4.2.1.5 GOLD

The Gaussian Of Local Descriptors (GOLD) [35] is based on a four-step algorithm: (i) evaluation of SIFT features; (ii) spatial pyramid decomposition; (iii) parametric probability density estimation; and (iv) projection of the covariance matrix onto the tangent Euclidean space in order to vectorize the feature.

4.2.1.6 HOG

The Histogram of Oriented Gradients (HOG) [36] groups pixels into small windows and measures intensity gradients in each of them. A histogram is then evaluated for each window, leading to the final descriptor. Windows of size 5×6 are used in our experiments.

4.2.1.7 AHP

The Adaptive Hybrid Pattern (AHP) [37] was created to overcome the two main drawbacks of the LBP feature, namely: its noisy behavior on quasi-uniform regions and its reactivity, i.e. the strong variations in the descriptor that are possibly induced by small variation in the input image (which is caused by the use of quantization thresholds).

4.2.1.8 FBSIF

Full BSIF [38] is an extension of the Binarized Statistical Image Feature (BSIF) [44] that assigns each pixel of the input image an n-bit label obtained by means of a set

of n linear filters. Each filter operates on a neighborhood of $l \times l$ pixels around the element that should be given the label. Such an n-bit label can be formalized as:

$$s = \boldsymbol{W}\boldsymbol{X},$$

where $\boldsymbol{X}$ is a vector of length $l^2 \times 1$ obtained from the neighborhood, while W is a $n \times l^2$ matrix that includes the filter vector notations. FBSIF operates by evaluating BSIF using several values of the filter size (SIZE_BSIF) and a binarization threshold (FULL_BSIF). Values considered in this work are: SIZE_BSIF $\in \{3, 5, 7, 9, 11\}$, FULL_BSIF $\in \{-9, -6, -3, 0, 3, 6, 9\}$. Each combination of size and threshold is fed to a separate SVM: the SVMs are then combined by sum rule.

4.2.1.9 COL

The color descriptor (COL) proposed in [39] is a simple and compact descriptor acquired by combining statistical measures extracted from each color channel in the RGB space. The final descriptor is obtained as the concatenation of several measures: the mean, the standard deviation, the 3rd and 5th moments of each color channel, and the marginal Histograms (8 bins per channel) [39].

4.2.1.10 MOR

The morphological descriptor (MOR) proposed in [40] is a set of measures extracted from a segmented version of the image, including aspect ratio, number of objects, area, perimeter, and eccentricity among others.

4.2.1.11 CLM

The CodebookLess Model (CLM) [41] is based on an image modeling method that can represent an image by means of a single Gaussian. This is obtained by first evaluating SIFT features on a regular grid placed on the image (this makes CLM a dense sampling features model), and then by fitting them using a Gaussian model. The main difference between CLM and the other widely used dense sampling method, such as the BoF approach [45], is the absence of a codebook.

In this work, according to the experiments reported in [26], we select as CLM the ensemble named CLoVo_3 in [26] based on e-SFT, PCA for dimensionality reduction, and one-vs-all SVM for the training phase.

LET

The LETRIST descriptor (LET) proposed in [42] is a simple but effective representation that explicitly encodes the joint information within an image across feature and scale spaces.

4.2.2 Deep Learned Features

4.2.2.1 Convolutional Neural Networks (CNN)

CNNs are built by repeated concatenation of five classes of layers: convolutional (CONV), activation (ACT), pooling (POOL), fully-connected (FC), and classification (CLASS) [46].

In our experiments we test and combine the following different CNN architectures:

- AlexNet [18]: this is the winner of the ImageNet ILSVRC challenge in 2012;
- GoogleNet [21]: this is the winner of the ImageNet ILSVRC challenge in 2014;
- VGGNet [20]: this network placed second in ILSVRC 2014, and the two best-performing VGG models (i.e. VGG-16 and VGG-19, with 16 and 19 weight layers) are available as pretrained models;
- ResNet [22]: this is the winner of ILSVRC 2015, and it is a network about 20 times deeper than AlexNet and 8 times deeper than VGGNet; we have tested ResNet50 and ResNet101;
- Inception [21]: this is a variant of GoogleNet that is based on the factorization of 7×7 convolutions into two or three consecutive layers of 3×3 convolutions;
- IncResv2 [47]: this is an Inception style network that makes use of residual connections instead of filter concatenation;
- DenseNet201 [48]: this is a very recent approach that connects each layer to every other layer of the network; for each layer, the inputs are all previous layers.

We test different values for the batch size and the learning rate:

- Batch size {10, 30, 50, 70};
- Learning rate {0.001, 0.0001}.

4.2.2.2 Fine-Tuning

A CNN is fine-tuned by retraining a pre-trained network to learn a different classification problem. Being aware of this problem, we adopt two different strategies for the fine-tuning of CNN:

- One round tuning (1R): where each pre-trained CNN and all its layers have been fine-tuned using the training set of the target problem;

- Two rounds tuning (2R): where the first round of tuning is performed by training a CNN using a leave-one-out dataset strategy, i.e. by including in the training set all the images from the datasets summarized in Sect. 4.2.2.1 except the training set that is the current target. The final number of classes is the sum of all the classes taken from each classification problem. The second round of tuning is the same as in "one round tuning" and involve only the training set of the target problem.

Datasets for 2R

Given the rationale of the Data Augmentation step, we use the following datasets in the first round of tuning:

- PAP: this is a PAP-SMEAR dataset [49]. PAP contains 917 images acquired during Pap tests and is used for cervical cancer diagnosis (available at http://labs.fme.aegean.gr/decision/downloads);
- LG: this is a "Liver gender" dataset [50]. LG includes 265 images of liver tissue sections from six-month male and female mice on a caloric restriction diet (the classes are the 2 genders);
- LA: this is a "Liver aging" [50] dataset. LA includes 529 images of liver tissue sections from female mice of four ages on an ad libitum diet;
- BR: this is a BREAST CANCER dataset [51]. BR contains 1394 images divided into the control, malignant cancer, and benign cancer classes;
- HI: this is a HISTOPATHOLOGY dataset [52]. HI contains 2828 images of connective, epithelial, muscular, and nervous tissue classes;
- RPE: this is a dataset composed of 195 human stem cell-derived retinal pigmented epithelium images that are divided into 16 subwindows, with each subwindow divided into four classes by two trained operators (available at https://figshare.com/articles/BioMediTech_RPE_dataset/2070109).

4.2.2.3 DeepScores (DS) [53]

This method performs transfer learning based on the analysis of neurons included in the deeper layers of the CNN. DS uses SVM classifiers by taking values directly from the low-level features, which are more general and depend only on the low-level patterns that were observed during the training of the CNN. Connecting these features to SVMs, however, is difficult to manage, since the typical output vector of a deep layer has a dimension that can be in the order of 10,000 elements and multiples. Dimensions of this size cannot be managed by an SVM (attempting to train such a large number would result in the curse of dimensionality). The use of dimensionality reduction methods before providing the data to the classifier is thus imperative.

In our experiments, we used two methods of dimensionality reduction that are widely adopted in the literature: Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT). Both were set to obtain a target descriptor length of 4000

Table 4.2 Layers used in each model for the DS ensemble. (note: we numbered all the layers of each network consecutively, the convolutional, ReLu, max-pooling layers, etc., and not separately according to type. Thus, layer 15 of AlexNet, for example, is the relu5 layer)

MODEL	LAYERS
AlexNet	15 17 20 23 24
GoogleNet	127 142 143
Vgg16	29 33 36 37 40
Vgg19	31 39 42 45
ResNet50	124 175 176
ResNet101	241 345 346
Inception	216 314 315

elements. In every case where the size of the feature vectors extracted from the deeper layers of the CNN exceeded 5000, PCA and DCT were used to reduce the dimension; otherwise the original feature vector was used as the input vector. We test SVM with histogram and radial basis function kernel as in [53]. All SVMs are then combined by sum rule (where for each tested layer, a different SVM is trained). Note that features are extracted using CNN after the tuning.

In Table 4.2, we report the layers used for describing an image. We choose layers similar to those used in [53], coupled with all the fully connected layers.

4.3 Results

4.3.1 Materials

We used the following two cervical cell datasets:

- **PAP**: this is the PAP SMEAR [49] dataset described above.
- **HEM**: this is the HEMLBC Dataset proposed in [54]. HEM contains cervical 989 abnormal cells from eight biopsy-confirmed CIN slides and 1381 normal cervical cells from another eight NILM (negative for intraepithelial lesion and malignancy) slides. For balancing data distribution, 989 normal cells are randomly selected. A 4-fold cross-validation was used where all the images of a given patient belong either to the training or to the testing set).

Moreover, we test our approach on the following medical image datasets:

- **BGR**: this is BREAST GRADING CARCINOMA [55] dataset of medium size containing 300 images (Grade 1: 107, Grade 2: 102 and Grade 3: 91 images) with a resolution of 1280×960 corresponding to 21 different patients with invasive ductal carcinoma of the breast. A 3-fold cross-validation is used on this dataset.
- **LY**: this is a LYMPHOMA dataset. LY includes 375 images of malignant lymphoma subdivided into three classes: CLL (chronic lymphocytic leukemia), FL (follicular lymphoma), and MCL (mantle cell lymphoma).

Table 4.3 Descriptive summary of the datasets

Dataset	#C	#S	URL for download
PAP	2	917	http://labs.fme.aegean.gr/decision/downloads
HEM	2	1978	Available upon request
LY	3	375	https://ome.grc.nia.nih.gov/iicbu2008/
BGR	3	300	https://zenodo.org/record/834910# .Wp1bQ-jOWUl
LAR	3	1320	https://zenodo.org/record/1003200# .WdeQcnBx0nQ
CO	8	5000	zenodo.org/record/53169# .WaXjW8hJaUm

- **LAR**: this is a LARYNGEAL dataset [56] that contains a well-balanced set of 1320 patches extracted from the endoscopic videos of 33 patients affected by SCC. The patches are relative to four laryngeal tissue classes.
- **CO**: this is a COLORECTAL dataset [57]. CO is a collection of textures obtained by manual annotation and tessellated of histological images of human colorectal cancer.

Table 4.3 summarizes each dataset in terms of the number of classes (#C), the number of samples (#S), and the URL for downloading the data. The testing protocol used in our experiments is the five-fold cross-validation method (except when the database is provided with its own protocol) and maintains the distribution of the classes in the datasets.

The experimental results reported in this section are based on two tests. In test 1, we measure the performance of handcrafted descriptors and their ensembles, without considering the DL-based methods. In test 2, DL methods are considered and fused with the handcrafted features. Ensembles are based on the score of each classifier after a normalization to 0-mean and standard deviation equal to 1; fusion is then performed by sum rule in all cases, which avoids any bias and tuning on a specific dataset and application scenario. The tests are performed on the pool of datasets mentioned above: the different nature of the images in the datasets provides a strong test for measuring how general the proposed approach is. Results, in turn, are compared against top approaches presented in the recent literature, again demonstrating the power of the general approach. The Wilcoxon signed-rank test was used to provide a statistical validation of the tests reported here.

The results of experiment 1 are available in Table 4.4, where accuracy is reported for the handcrafted descriptors and their best-performing combinations, tested on the six datasets considered. In detail, FH is the fusion of LTP, CLBP, RIC, LET, MOR, AHP, COL, MLPQ and FullBSIF, namely, all the handcrafted methods, with the exception of GOLD and CLM, which are more computationally expensive than the others. The performance gain provided by these two demanding features is separately analyzed and reported as FH+CLM and FH+CLM+GOLD. It should be noted that MLPQ and FullBSIF are based on ensembles of features, but they have the same weight as the other single features in this fusion mechanism. Unsurprisingly, the best performance is achieved using the full set of features, with the p-value 0.05.

Table 4.4 Handcrafted descriptors

	PAP	HEM	LY	CO	BGR	LAR
LTP	87.14	82.13	85.33	90.40	87.54	71.97
MLPQ	88.12	82.44	92.27	93.58	90.54	82.27
CLBP	85.61	82.45	86.67	92.04	89.54	72.27
RIC	90.73	82.54	85.87	91.56	91.87	90.68
LET	85.49	77.18	92.53	93.18	93.54	90.76
MOR	91.93	78.86	84.53	93.30	91.54	79.85
AHP	92.48	84.39	93.87	94.16	91.37	85.30
COL	88.88	87.61	91.47	92.30	90.71	85.30
FBSIF	88.01	85.17	92.53	93.42	88.00	88.56
GOLD	87.89	84.08	53.07	83.58	75.33	90.61
CLM	88.11	82.65	74.40	89.60	86.33	87.58
FH	90.85	86.08	95.20	95.18	91.67	91.29
FH+CLM	90.41	85.81	94.93	95.08	91.67	92.12
FH+CLM+GOLD	89.97	86.37	93.60	94.92	92.00	93.26

The results of the second experiment are reported in Table 4.5 (for standard tuning) and Table 4.6 (for 2-round tuning, as described in Sect. 4.2). Tests were performed by tuning the networks multiple times varying two learning parameters, namely:

- Learning Rate (LR) taking the values 10^{-3} and 10^{-4};
- Batch Size (BS) taking values in {10, 30, 50, 70}.

Each network was characterized on all the datasets considered. For each dataset, we report the performance provided by:

- The Single Best (SB) CNN, i.e. the CNN that provided the best performance for that specific dataset; a certain degree of overfitting is present in this case, but this value is used as a baseline to compare the ensembles;
- All Best (AB) CNN, which is the result obtained for that specific dataset provided by the network that provided the best average performance on all the datasets;
- The Fusion (Fus) of all the CNNs trained on all the combinations of BS and LR. CNNs that did not converge (e.g. AlexNet and VGG using LR $= 10^{-4}$) or exceeded the memory available on the GPU were excluded from the ensemble;
- The FCN-ST (xxx-Standard Tuning), representing the fusion of all the methods in Fus (using standard tuning), again excluding the networks that were not successfully tuned. In this case, the number of successful tunings is taken into account: the score assigned to a network type, say, GoogleNet, is evaluated as the average of all the GoogleNets that were successfully tuned (but with different parameters). Once the score of all network types is evaluated, they are fused together. Two versions of FCN-st are reported in Table 4.5: "All" and "NoAlex" (in the latter case, AlexNet is excluded from the ensemble due to its low performance);

Table 4.5 Standard tuning

	AlexNet			VGG16			VGG19			IncResv2			FCN-st	
	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	All	NoAlex
PAP	90.19	90.19	90.19	91.83	91.83	92.04	91.39	91.39	91.71	92.16	92.16	91.49	93.67	94.11
HEM	87.06	82.40	86.17	86.43	83.63	85.85	87.45	87.45	87.66	85.98	85.98	86.99	**88.60**	88.45
LY	82.40	82.40	79.73	80.80	80.80	85.33	82.40	82.40	86.13	84.80	84.80	85.87	93.87	**94.40**
CO	94.22	94.22	95.14	96.14	96.14	96.90	95.94	95.26	96.76	93.58	93.58	95.16	97.38	**97.40**
BGR	92.00	91.00	91.33	93.00	93.00	95.00	93.67	93.67	91.67	91.00	91.00	90.67	**96.00**	95.67
LAR	90.68	89.39	90.08	93.33	91.52	91.97	94.24	93.26	**95.38**	94.62	94.62	94.39	94.70	94.77
AVG	89.42	88.26	88.77	90.25	89.48	91.18	90.84	90.57	91.55	90.35	90.35	90.76	94.03	**94.13**

	GoogleNet			ResNet50			ResNet101			Inception			DenseNet		
	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus
PAP	91.72	91.28	91.17	92.70	92.70	92.91	90.84	90.84	92.15	92.71	92.71	93.68	94.23	94.23	**95.10**
HEM	87.82	86.66	87.69	86.59	86.59	88.04	84.70	82.48	85.59	88.03	87.07	87.90	88.71	88.25	88.35
LY	82.93	82.93	83.73	86.40	86.40	89.87	86.40	86.40	86.13	87.47	87.47	89.60	86.93	86.93	89.07
CO	95.60	95.60	96.30	95.42	95.42	96.40	92.92	92.92	94.68	95.02	92.82	96.40	96.68	95.24	97.14
BGR	93.00	90.83	94.67	91.33	90.33	93.67	93.33	93.33	93.33	93.67	93.67	94.67	91.00	91.00	92.33
LAR	92.35	90.83	91.97	92.20	92.05	93.26	93.64	93.64	93.86	92.73	89.77	93.48	94.02	93.64	94.39
AVG	90.57	89.68	90.92	90.77	90.58	92.35	90.30	89.93	90.95	91.60	90.58	92.62	91.92	91.54	92.73

Table 4.6 Two-round tuning

	AlexNet			GoogleNet			VGG16			VGG19			FCN-Two	
	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	SB	AB	Fus	All	NoAlex
PAP	92.81	91.82	93.46	93.13	91.94	93.68	93.37	91.49	**94.88**	93.24	90.43	94.22	94.45	94.77
HEM	88.44	88.44	88.83	87.35	87.35	87.91	88.81	87.08	89.26	**90.34**	88.29	90.12	90.14	89.67
LY	86.93	86.93	86.93	85.87	85.87	87.47	85.33	85.33	86.13	89.33	89.33	90.67	**95.47**	94.67
CO	94.70	92.48	95.48	95.48	95.28	96.50	96.16	95.88	97.04	96.62	95.62	**97.44**	97.20	97.34
BGR	91.67	90.33	92.00	92.33	92.33	93.00	92.23	92.00	94.33	92.00	92.00	94.00	94.33	**95.33**
LAR	92.42	90.23	92.05	91.29	90.00	92.12	94.02	94.02	93.26	93.48	93.48	**95.08**	94.24	94.70
AVG	91.16	90.03	91.45	90.90	90.46	91.78	91.65	90.96	92.48	92.50	91.52	93.58	94.30	**94.41**

- FCN-two: is the same as FCN-ST, but uses two-round tuning instead of standard tuning.

The last row in Tables 4.5 and 4.6 report the average performance in the tested datasets.

Results reported in Tables 4.5 and 4.6 show:

- For each topology, Fus outperforms AB (Wilcoxon signed-rank test—p-value of 0.05);
- The best CNN is DenseNet for standard tuning and VGG19 for two-round tuning, but it should be noted that, due to computational issues for two-round tuning, we run only four CNNs.
- FCN-st outperforms each Fus—Standard Tuning (Wilcoxon signed-rank test—p-value of 0.05).
- FCN-two obtains a performance that is similar to FCN-st (which is true when we used only 4 CNNs in FCN-two).

In Table 4.7, we report the performance of the transfer learning approach detailed in Sect. 4.2.2.3. Due to computational issues, we run tests only on PAP and HEM. We compare DS with other, more standard, approaches where the last layers of CNN are used to train SVM without any dimensionality reduction (reported in Table 4.8). Notice that for some CNNs only one layer can be used for feeding an SVM without any dimensionality reduction. The reported results are obtained by combining, via sum rule, all the SVMs trained with all the CNNs.

Clearly, SVM produces excellent results when trained with CNN features, and the performance improves when more layers are considered. The performance improvement of DS with respect to AF is negligible. More datasets should be used for a more reliable comparison of DS with respect to AF. In DS, LF, LT, and AF, we use only

Table 4.7 SVMs trained with CNN features

	DS+	DS	LF	LT	AF
PAP	**94.33**	93.90	92.25	92.79	93.45
HEM	**89.10**	88.48	88.13	88.78	88.38

Table 4.8 Layers used for feeding SVM without dimensionality reduction

MODEL	LF	LT	AF
AlexNet	23	20 23	17 20 23
GoogleNet	142	142	142
Vgg16	39	36 39	33 36 39
Vgg19	45	42 45	39 42 45
ResNet50	175	175	175
ResNet101	345	345	345
Inception	314	314	314

Table 4.9 Ensembles proposed here

	FCN+		Here1	Here2	Here1+	Here2+
	All	NoAlex				
PAP	94.66	94.55	94.33	94.33	**94.88**	94.55
HEM	89.28	89.02	**89.48**	89.33	89.37	89.30
LY	94.67	94.93	**97.33**	96.53	–	–
CO	97.23	**97.50**	97.26	97.20	–	–
BGR	**96.33**	96.00	95.33	95.33	–	–
LAR	94.77	94.85	95.38	**95.45**	–	–

the CNN coupled with the standard tuning. In DS+ we report the performance of the method DS coupled with both the standard tuning and two-round tuning.

In Table 4.9, the ensemble of CNNs is combined with other methods. The ensembles evaluated are the following:

- FCN+ : sum rule among the methods that belong to FCN-st and FCN-Two;
- Here1: sum rule between (FCN+ NoAlex) and FH, before fusion the scores of (FCN+ NoAlex) and FH are normalized to mean 0 and standard deviation 1.
- Here2: sum rule between (FCN+ NoAlex) and (FH+ CLM+GOLD), before fusion the scores of (FCN+ NoAlex) and FH are normalized to mean 0 and standard deviation 1.
- Here1+ : sum rule among (FCN+ NoAlex), FH and DS+ , before fusion the scores of each method are normalized to mean 0 and standard deviation 1.
- Here2+ : sum rule among (FCN+ NoAlex), (FH+CLM+GOLD) and DS+ , before fusion the scores of each method are normalized to mean 0 and standard deviation 1.

Both Here1 and Here2 outperform FCN+; hence, the handcrafted features are useful for improving the performance of an ensemble of CNNs. Since Here1 is simpler then Here2, our suggestion is to use Here1.

4.4 Conclusion

In this work, a set of descriptors based on learned and handcrafted features has been proposed. The extracted features are based on different approaches: local features; dense sampling features; and deep learning. Each descriptor is used to train a different classifier; in this way an ensemble of classifiers is obtained. The final score of a given test pattern is calculated by sum rule. We investigated deep learning approaches based on different methods of training: fine-tuning of pre-trained network and transfer learning from different levels of the networks fed to SVM classifiers. Several datasets were used to validate our approach further, and the results clearly show that the proposed ensemble obtains state-of-the-art performance and is, thus, generalizable.

In the future, we plan on exploring different transfer learning approaches based on tuned CNNs.

The MATLAB code used in this chapter will be available at https://github.com/LorisNanni.

Acknowledgements We gratefully acknowledge the support of: NVIDIA Corporation "NVIDIA Hardware Donation Grant" with the donation of the Titan X used for this research; National Natural Science Foundation of China (81501545).

References

1. Zhou, J., Lamichhane, S., Sterne, G., Ye, B., Peng, H.: BIOCAT: a pattern recognition platform for customizable biological image classification and annotation. BMC Bioinform. **14**, 291 (2013)
2. Misselwitz, B., et al.: Enhanced CellClassifier: a multi-class classification tool for microscopy images. BMC Bioinform. **11**(30) (2010)
3. Pau, G., Fuchs, F., Sklyar, O., Boutros, M., Huber, W.: EBImage—an R package for image processing with applications to cellular phenotypes. Bioinformatics **26**(7), 979–981 (2010)
4. Uhlmann, V., Singh, S., Carpenter, A.E.: CP-CHARM: segmentation-free image classification made accessible. BMC Bioinform. **17**, 51 (2016)
5. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.J.: Image classification for content-based indexing. IEEE Trans. Image Process. **10**(1), 117–130 (2001)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Addison-Wesley Longman Publishing Co., Inc, Boston (2001)
7. Xu, Y., Huang, S., Ji, H., Fermüller, C.: Scale-space texture description on SIFT-like textons. Comput. Vis. Image Underst. **116**(9), 999–1013 (2012)
8. Ojala, T., Pietikainen, M., Maeenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
9. Nanni, L., Lumini, A., Brahnam, S.: Survey on LBP based texture descriptors for image classification. Expert Syst. Appl. **39**(3), 3634–3641 (2012)
10. Vu, T.H., Mousavi, H.S., Monga, V., Rao, G., Rao, A.: Histopathological image classification using discriminative feature-oriented dictionary learning. IEEE Trans. Med. Imaging **35**(3), 738–751 (2016)
11. Otalora, S., et al.: Combining unsupervised feature learning and riesz wavelets for histopathology image representation: application to identifying anaplastic medulloblastoma. Presented at the international conference on medical image computing and computer assisted intervention, Munich (2015)
12. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
13. Greenspan, H., van Ginneken, B., Summers, R.M.: Deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging **35**, 1153–1159 (2016)
14. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inform. **7**(29) (2016)
15. Gua, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018)
16. Russakovsky, O., Deng, J., Su, H.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)

17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2323 (1998)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 1097–1105. Curran Associates Inc, Red Hook, NY (2012)
19. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision—ECCV 2014. ECCV 2014, Lecture Notes in Computer Science, vol. 8689. Springer, Berlin, Cham (2014)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Cornell University (2014). arXiv:1409.1556v6
21. Szegedy, C., et al.: Going deeper with convolutions. Presented at the IEEE computer society conference on computer vision and pattern recognition (2015)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Presented at the 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV (2016)
23. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Cornell University. arXiv:1411.17922014
24. Shin, H.-C., et al.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
25. Pan, Y., et al.: Brain tumor grading based on neural networks and convolutional neural networks. Presented at the 37th IEEE engineering in medicine and biology society (EMBC) (2015)
26. Nanni, L., Brahnam, S., Ghidoni, S., Lumini, A.: Bioimage classification with handcrafted and learned features. IEEE/ACM Trans. Comput. Biol. Bioinf. **16**(3), 874–885 (2018)
27. van Ginneken, B., Setio, A.A.A., Jacobs, C., Ciompi, F.: Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. Presented at the IEEE 12th international symposium on biomedical imaging (ISBI) (2015)
28. Li, R., et al.: Deep learning based imaging data completion for improved brain disease diagnosis. Med. Image Comput. Comput.-Assist. Interv. **17**(Pt 3), 305–312 (2014)
29. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Health Inform. **21**(1), 31–40 (2017)
30. Nanni, L., Ghidoni, S., Brahnam, S.: Ensemble of convolutional neural networks for bioimage classification. In: Applied Computing and Informatics. In press
31. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: Analysis and Modelling of Faces and Gestures, LNCS, vol. 4778, pp. 168–182 (2007)
32. Nanni, L., Brahnam, S., Lumini, A.: A very high performing system to discriminate tissues in mammograms as benign and malignant. Expert Syst. Appl. **39**(2), 1968–1971 (2012)
33. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. **19**(6), 1657–1663 (2010)
34. Nosaka, R., Fukui, K.: HEp-2 cell classification using rotation invariant co-occurrence among local binary patterns. Pattern Recognit. Bioinform. **47**(7), 2428–2436 (2014)
35. Serra, G., Grana, C., Manfredi, M., Cucchiara, R.: Gold: Gaussians of local descriptors for image representation. Comput. Vis. Image Underst. **134**(May), 22–32 (2015)
36. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Presented at the 9th European conference on computer vision, San Diego, CA (2005)
37. Zhu, Z., et al.: An adaptive hybrid pattern for noise-robust texture analysis. Pattern Recogn. **48**, 2592–2608 (2015)
38. Nanni, L., Paci, M., Santos, F.L.C.D., Brahnam, S., Hyttinen, J.: Review on texture descriptors for image classification. In: Alexander, S. (ed.) Computer Vision and Simulation: Methods, Applications and Technology. Nova Publications, Hauppauge, NY (2016)

39. Bianconi, F., Fernández, A., González, E., Saetta, S.A.: Performance analysis of colour descriptors for parquet sorting. Expert. Syst. Appl. **40**(5), 1636–1644 (2013)
40. Strandmark, P., Ulén, J., Kahl, F.: HEp-2 staining pattern classification. Presented at the international conference on pattern recognition (ICPR2012) (2012). https://lup.lub.lu.se/search/ws/files/5709945/3437301.pdf
41. Wang, Q., Li, P., Zhang, L., Zuoc, W.: Towards effective codebookless model for image classification. Pattern Recogn. **59**, 63–71 (2016)
42. Song, T., Meng, F.: Letrist: locally encoded transform feature histogram for rotation-invariant texture classification. IEEE Trans. Circuits Syst. Video Technol. **PP**(99) (2017)
43. Nanni, L., Brahnam, S., Lumini, A., Barrier, T.: Ensemble of local phase quantization variants with ternary encoding. In: Brahnam, S., Jain, L.C., Lumini, A., Nanni, L. (eds.) Local Binary Patterns: New Variants and Applications, pp. 177–188. Springer, Berlin (2014)
44. Kannala, J., Rahtu, E.: Bsif: binarized statistical image features. Presented at the 21st international conference on pattern recognition (ICPR 2012), Tsukuba, Japan (2012)
45. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. Presented at the European conference on computer vision (ECCV) (2006)
46. Goodfellow, A., Ian, B., Yoshua, C.: Deep Learning. MIT Press (2016)
47. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. "arxiv.org," Cornell University. https://arxiv.org/pdf/1602.07261.pdf2016
48. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. CVPR **1**(2), 3 (2017)
49. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: Pap-smear benchmark data for pattern classification. Presented at the nature inspired smart information systems (NiSIS), Albufeira, Portugal (2005)
50. Shamir, L., Orlov, N.V., Eckley, D.M., Goldberg, I.: IICBU 2008: a proposed benchmark suite for biological image analysis. Med. Biol. Eng. Compu. **46**(9), 943–947 (2008)
51. Junior, G.B., Cardoso de Paiva, A., Silva, A.C., Muniz de Oliveira, A.C.: Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM. Comput. Biol. Med. **39**(12), 1063–1072 (2009)
52. Cruz-Roa, A., Caicedo, J.C., González, F.A.: Visual pattern mining in histology image collections using bag of features. Artif. Intell. Med. **52**, 91–106 (2011)
53. Nanni, L., Ghidoni, S., Brahnam, S.: Handcrafted vs non-handcrafted features for computer vision classification. Pattern Recogn. **71**, 158–172 (2017)
54. Xhang, L., Lu, L., Nogues, I.: Deeppap: deep convolutional networks for cervical cell classification. IEEE J. Biomed. Health Inform. **21**(6) (2017)
55. Dimitropoulos, K., Barmpoutis, P., Zioga, C., Kamas, A., Patsiaoura, K., Grammalidis, N.: Grading of invasive breast carcinoma through Grassmannian VLAD encoding. PLoS ONE **12**, 1–18 (2017)
56. Moccia, S., et al.: Confident texture-based laryngeal tissue classification for early stage diagnosis support. J. Med. Imaging (Bellingham) **4**(3), 34502 (2017)
57. Kather, J.N., et al.: Multi-class texture analysis in colorectal cancer histology. Sci. Rep. **6**, 27988 (2016)

Chapter 5
Deep Unsupervised Representation Learning for Audio-Based Medical Applications

Shahin Amiriparian, Maximilian Schmitt, Sandra Ottl, Maurice Gerczuk and Björn Schuller

Abstract Feature learning denotes a set of approaches for transforming raw input data into representations that can be effectively utilised in solving machine learning problems. Classifiers or regressors require training data which is computationally suitable to process. However, real-world data, e.g., an audio recording from a group of people talking in a park whilst in the background a dog is barking and a musician is playing the guitar, or health-related data such as coughing and sneezing recorded by consumer smartphones, comprises a remarkably variable and complex nature. For understanding such data, developing expert-designed, hand-crafted features often demands for an exhaustive amount of time and resources. Another disadvantage of such features is the lack of generalisation, i.e., there is a need for re-engineering new features for new tasks. Therefore, it is inevitable to develop automatic representation learning methods. In this chapter, we first discuss the preliminaries of contemporary representation learning techniques for computer audition tasks. Hereby, we differentiate between approaches based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). We then introduce and evaluate three state-of-the-art deep learning systems for unsupervised representation learning from raw audio: (1) pre-trained image classification CNNs, (2) a deep convolutional generative adversarial network (DCGAN), and (3) a recurrent sequence-to-sequence autoencoder (S2SAE). For each of these algorithms, the representations are obtained from the spectrograms of the input audio data. Finally, for a range of audio-based machine learning tasks, including abnormal heart sound classification, snore sound

S. Amiriparian (✉) · M. Schmitt · S. Ottl · M. Gerczuk · B. Schuller
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
e-mail: amiriparian@ieee.org

B. Schuller
e-mail: schuller@ieee.org

B. Schuller
GLAM – Group on Language, Audio and Music, Imperial College London, London, UK

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_5

classification, and bipolar disorder recognition, we evaluate the efficacy of the deep representations, which are: (i) the activations of the fully connected layers of the pre-trained CNNs, (ii) the activations of the discriminator in case of the DCGAN, and (iii) the activations of a fully connected layer between the encoder and decoder units in case of the S2SAE.

5.1 Background

Before explaining how the methodologies described in Sect. 5.2 can be used for the purpose of unsupervised representation learning in detail, the following section will outline the four neural network architectures and models that form their basis. In particular, two types of neural networks—Convolutional Neural Networks and Recurrent Neural Networks—and two higher level models—Autoencoders and Generative Adversarial Neural Networks—will be discussed.

5.1.1 Convolutional Neural Networks

The first neural network architecture we consider for representation learning are *Convolutional Neural Networks* (*CNNs*). CNNs rose to popularity in the machine learning community for their efficacy of solving visual recognition tasks, like image [1–3] and video classification [4] and action [5, 6] or face recognition/detection [7–9]. Classic CNNs make use of two distinct types of hidden layers: the eponymous *convolutional* layers and *pooling* layers. Convolutional layers convolve their inputs with striding kernels of a specific, small size, resulting in so-called feature maps. The parameters of each kernel are shared between convolutions with different parts of the input. As a result, the network is able to recognise different features of the input, like detecting edges, regardless of where they are located in the original image. On the other hand, pooling layers reduce the number of neurons of feature maps by putting together groups of adjacent neurons. One of the most popular pooling approaches is to employ *max-pooling*: small square regions of the feature maps are reduced by the maximum activation in this region. Pooling leads to a smaller number of trainable parameters and can also help to prevent overfitting and improve generalisation capabilities of the network [10]. Conventional fully connected layers are also often added after these two special layers and, in the case of classification tasks, a softmax layer can be used to complete the architecture. Figure 5.1 shows a simple CNN architecture.

Apart from the field of computer vision, CNNs have also been successfully applied in other research areas. In Natural Language Processing they have been used for tasks like sentence classification [11, 12] or sentiment analysis [13, 14], whilst they also achieve state-of-the-art results in computer audition for example in environmental sound [15, 16] or acoustic scene classification [17, 18].

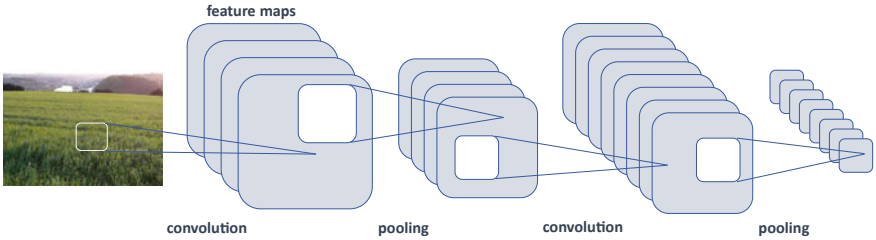


Fig. 5.1 Architecture of a simple Convolutional Neural Network: kernels are convolved with the input to create feature maps in the convolutional layers. Pooling is often used as a dimensionality reduction technique afterwards

5.1.2 Generative Adversarial Networks

Another technique of unsupervised representation learning which we evaluate for the task of audio analysis are Generative Adversarial Networks (GANs). In GANs, two different networks are trained simultaneously in a zero-sum game: a so called *Generator* creates samples from a random distribution, typically a noise vector z, whereas a *Discriminator* is trained to distinguish these generated samples from real data. This adversarial setting leads to the need of both models to continually increase their performance until the samples produced by the generator become indistinguishable from the real data distribution [19] or at least an equilibrium is reached, in which both models cannot improve further.

Chen et al. extend GANs from an information-theoretic point of view with Information Maximising Generative Adversarial Networks (InfoGANs) [20]. InfoGANs learn interpretable representations in an unsupervised manner by maximising mutual information between a set of latent variables and the generator distribution. The noise vector is split up into a source of incompressible noise z and a set of latent variables, denoted as latent code c. This code targets structured salient information of the data distribution, e.g., for the MNIST database of handwritten digits [21] individual variables of c learn to represent the digit kind, the width or the rotation of the generated character.

Wasserstein-GANs [22] minimise an efficient approximation of the Earth Mover distance between generated and real data distribution to effectively combat a main problem of training GANs: Normally, GANs require balanced training of generator and discriminator and are quite sensitive to changes in the network architecture [23].

Whilst the samples generated by a GAN are often indistinguishable from the real data distribution by the human eye, Valle et al. recently showed that these fake samples carry a unique signature that makes them easily identifiable using methods of statistical analysis and pixel value comparison [24]. The samples also violate formal specifications that can be learnt from the respective real data (Fig. 5.2).

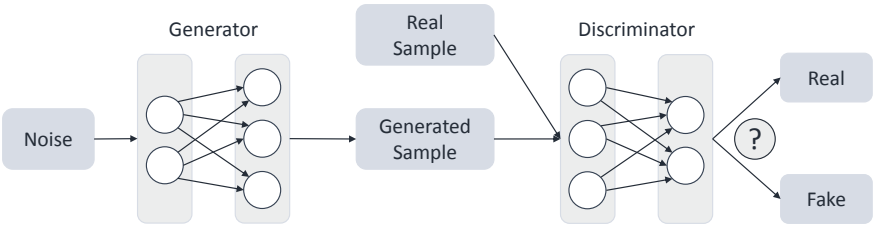


Fig. 5.2 Basic concept of a Generative Adversarial Network: the *generator* constructs samples from noise that should resemble samples from the real data distribution. The *discriminator* is then tasked with telling these "fake" examples from real ones

5.1.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of neural network that aims at modelling time-series data. A standard RNN is similar in structure to a simple multi-layer perceptron model, but introduces connections between hidden layers of different time steps. These connections can be of different form (cf. Fig. 5.3): for one, neurons can be connected to themselves between time steps (direct). Neurons can also be connected to neurons on the same (lateral) or different (indirect) hidden layer. This enables the network to "learn from the past", i.e., discovering temporal correlations in the data distribution. Training RNNs can be achieved by backpropagation through time (BPTT): the recurrent network is "unrolled" into a multilayer network consisting of copies of the network for each time step. Regular RNNs are comparably difficult to train, as the BPTT algorithm leads to some problems: Correlating similarities between data points separated by longer periods of time is hindered by *vanishing* and *exploding* gradients [25–27]. A RNN architecture designed specifically to tackle

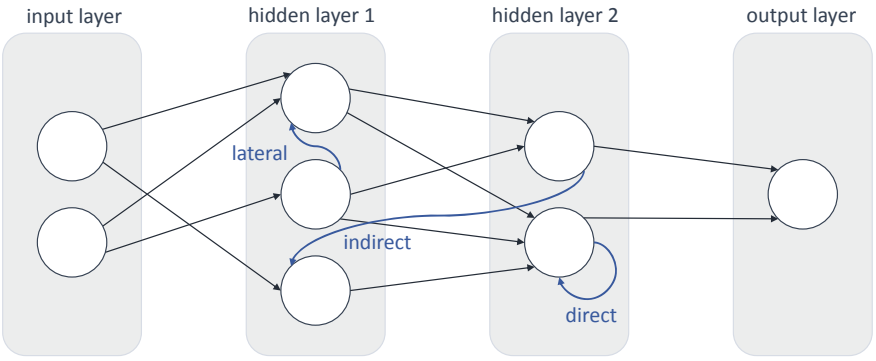


Fig. 5.3 An RNN architecture showing three different types of recurrent connections (blue arrows). *Direct* connections link a neuron to itself between time stamps. Connections between different neurons are called *lateral* when the neurons reside on the same hidden layer and *indirect* if they are on separate layers

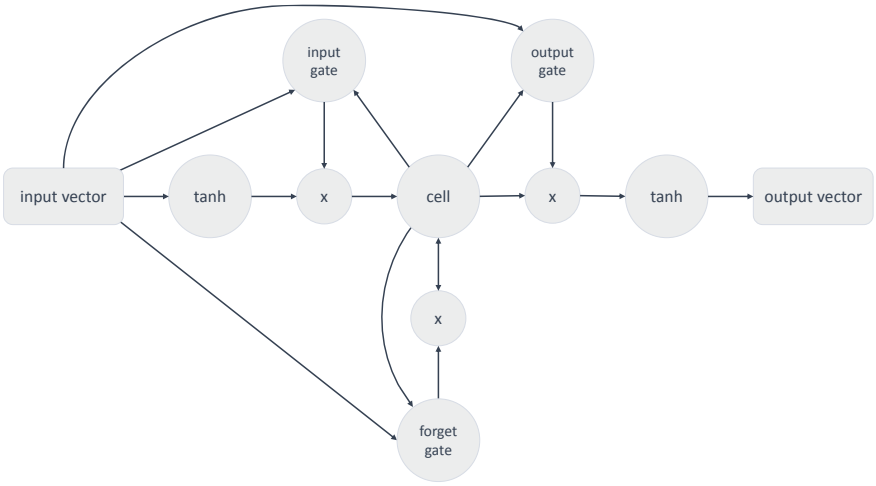


Fig. 5.4 Basic building block of the *Long Short-Term Memory architecture*: a memory *cell* stores a hidden state, whereas different *gate units* regulate how this state should be affected by the input (*input gate*), influence the output (*output gate*) or even be forgotten after certain events (*forget gate*)

this problem is the so-called Long Short-Term Memory (LSTM) network [28, 29]. LSTMs enforce constant error flow through the use of memory cells in combination with gate units. A memory cell stores a hidden state, while multiplicative gate units control how this state is influenced by the current input (*input gate*) as well as how it should affect the current output (*output gate*). Figure 5.4 shows this architectural concept extended with *forget gates* which were introduced to help LSTMs handle very long or continuous sequences without pre-defined beginning and end. These gates enable the network to reset its internal state at a certain time.

5.1.4 Autoencoders

Autoencoders are neural networks that learn compressed, efficient data coding by unsupervised training. The encoder part of the network maps input data to a lower dimensional representation, e.g., by stacking fully connected layers that decrease in neuron count. The decoder part of the network then tries to reconstruct the input data from this compressed representation. The training objective of the network is therefore to minimise the difference (measured for example by mean squared error) between the original data it receives as input and the output it reconstructs from the learnt coding. The simplest form of such a network would be a multilayer perceptron architecture having an input and output layer of the same size and hidden layers in between which first contract in size up to a specific layer which represents the learnt coding and then expands again before the output layer (cf. Fig. 5.5).

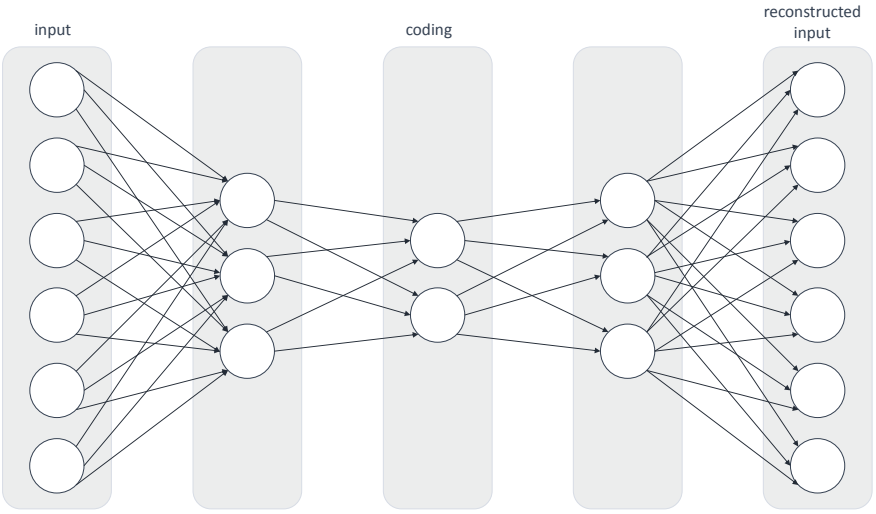


Fig. 5.5 A simple autoencoder architecture: the network learns a compressed representation of the input, the so-called *coding*, from which it can still accurately reconstruct the original input

Variations of autoencoders exist that aim to learn richer representations and prevent the networks from learning the identity function. Noteworthy are for example denoising autoencoders. Here, the input data is intentionally corrupted before being fed to the encoder network. The autoencoder is then trained to reconstruct the original, clean data from the corrupted samples [30].

5.2 Deep Representation Learning Methodologies

We now analyse three state-of-the-art deep learning methodologies for unsupervised representation learning from the acoustic sounds reflecting our physiological and pathological states introduced in Sect. 5.3. These techniques are based on the neural networks approaches discussed in Sect. 5.1 and are developed to cope with the data scarcity and representation learning challenges which are common issues in machine learning tasks, especially for medical applications [31–33]. As a possible solution for data scarcity problems in the audio domain, we introduce a transfer learning approach, namely the DEEP SPECTRUM system (cf. Sect. 5.2.1), which uses CNNs pre-trained on ImageNet [34] data for extracting features from the audio modality. In Sect. 5.2.2, we introduce a deep convolutional generative adversarial network (DCGAN) for unsupervised learning of robust representations from spectral features, and a recurrent sequence to sequence autoencoder for learning fixed-length representations of variable-length audio sequences is proposed in Sect. 5.2.3 (Fig. 5.6).

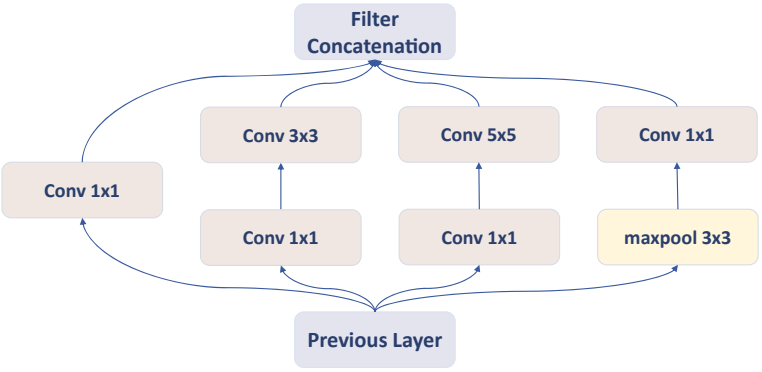


Fig. 5.6 An inception module used in the GoogLeNet architecture. Small 1×1 convolutions are applied to reduce the dimensionality. To combine information found at different scales, filters of different path sizes are concatenated. Figure adapted from [45]

5.2.1 Pre-trained Convolutional Neural Networks

To solve complex recognition tasks, machine learning systems should have a considerable amount of prior knowledge to compensate for the data which is not available. Convolutional neural networks (CNNs) are able to establish one such class of models [1, 35–38]. Hence, it is possible to transfer the knowledge of deep CNNs which have been pre-trained on large-scale image data (e.g., ImageNet) to other (computer vision) tasks in which the data is scarce, such as medical image analysis [39, 40] or text classification [12, 41]. The process of transferring of the stored knowledge from one domain to another domain is defined as transfer learning [42]. In this chapter, we analyse the efficacy of such an approach for audio-based recognition tasks by introducing the DEEP SPECTRUM system[1] [43], which utilises pre-trained CNNs, including AlexNet [1], VGG16 and VGG19 [3], and GoogLeNet [44]. All four architectures have been trained for the task of object recognition on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data.[2]

5.2.1.1 Deep Spectrum System

An overview of the DEEP SPECTRUM system is given in Fig. 5.7. In the pre-processing step, two-dimensional visual representations of the input audio files, such as spectrograms or mel-spectrograms are created (cf. Sect. 5.2.1.2). This step is necessary, as the CNN descriptors use two-dimensional filters to process the input spectral representations (cf. Sect. 5.2.1.3). After forwarding the spectrograms through the CNNs, the activations of the fully connected layers of each pre-trained network are

[1] https://github.com/DeepSpectrum/DeepSpectrum.
[2] http://www.image-net.org/challenges/LSVRC.

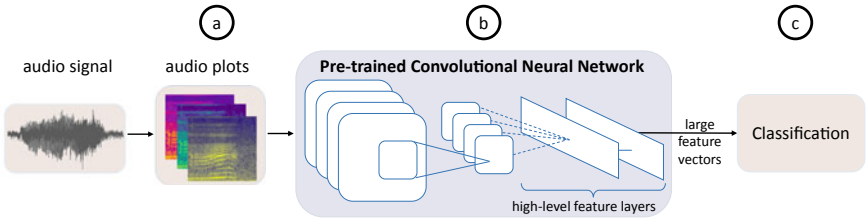


Fig. 5.7 Illustration of the DEEP SPECTRUM system. Pre-trained CNNs are used to obtain task-dependent representations from the audio signals

extracted as feature vectors. These high-level, shift-invariant features, denoted as DEEP SPECTRUM features, are used to train a classifier (cf. Fig. 5.7c). It is worth mentioning that the convolutional layers are able to make strong assumptions about the locality of pixel dependencies, i.e., the more local structures are available in the generated spectrograms, the more robust are the extracted DEEP SPECTRUM features.

5.2.1.2 Mel-spectrogram Creation

The mel-spectrograms (cf. Fig. 5.7a) are calculated with a window size of w and an overlap of $0.5w$ from the log-magnitude spectrum by dimensionality reduction using a mel-filter with N_{mel} filter banks equally distributed on the mel-scale defined in Eq. (5.1):

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right), \tag{5.1}$$

where f_{mel} is the resulting frequency on the mel-scale computed in mels and f_{Hz} is the normal frequency measured in Hz. The mel-scale is based on the frequency response of the human ear that has better resolution at lower frequencies. We also display the mel-spectrogram on this scale. A sample mel-spectrogram plot for a member of the *Snore* class (cf. Sect. 5.3.2) is shown in Fig. 5.8. It has been demonstrated that different DEEP SPECTRUM feature vectors will be extracted when the colour maps are changed for the same input images [43, 45, 46]. Based on the results in [43, 45, 47, 48], the DEEP SPECTRUM representations extracted from the audio plots with *viridis* colour map show stronger performance than a range of other colour maps, such as *cividis*, *hot*, *magma*, *plasma*, or *vega20b*. We assume that the main reason for this effect is the capability of *viridis*—which is a perceptually uniform sequential colour map varying from blue (low range) to green (mid range) to yellow (upper range) (cf. Fig. 5.8)—to cover a wide range of the colours available in the ImageNet pictures.

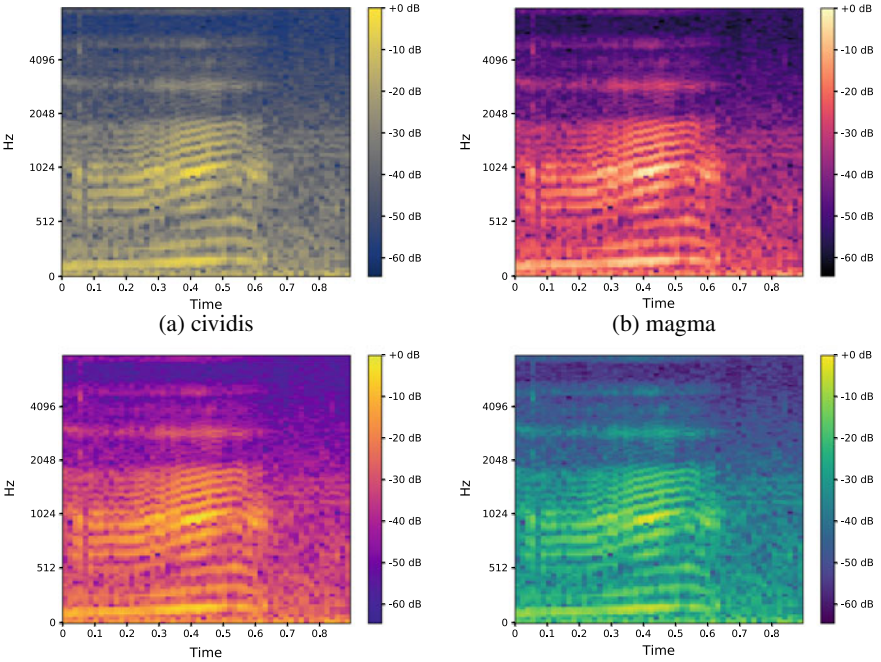


Fig. 5.8 A mel-spectrogram plot of an Snoring audio sample (file id: train_0261.wav) from the MPSSC dataset [49] with four different colour maps, *cividis*, *magma*, *plasma*, and *viridis*. The colour bar to the right shows the colour changes associated with increasing spectral energy

5.2.1.3 Applied Pre-trained CNNs

To form suitable deep representations from the mel-spectrograms, four pre-trained CNNs, AlexNet, VGG16 and VGG19, and GoogLeNet, are applied as feature extractors. The architectures of AlexNet, and VGG networks are compared in Table 5.1. The inception module used in GoogLeNet is shown in Fig. 5.6.

AlexNet

AlexNet has 5 convolutional layers, in cascade with 3 fully connected layers [1]. An overlapping max-pooling operation is applied to downsample the feature maps generated by the first, second, and third convolutional layers. A rectified linear unit (ReLU) non-linearity is used, as this non-saturating function regularises the training, whilst improving the network's generalisation capabilities. The fully connected layers *fc6*, *fc7*, and *fc8*, have 4096, 4096, and 1000 neurons, respectively. We use the activations of the neurons in the fully connected layers as feature vectors for our experiments in Sect. 5.3.

Table 5.1 Comparison between three pre-trained CNNs, AlexNet, VGG16, and VGG19, for extracting DEEP SPECTRUM features. *ch* stands for channels and *conv* denotes convolutional layers. The table is adapted from [43]

AlexNet	VGG16	VGG19
Input: RGB image		
1×conv size: 11; ch: 96; stride: 4	2×conv size: 3; ch: 64; stride: 1	
Maxpooling		
1×conv size: 5; ch: 256	2×conv size: 3; ch: 128	
Maxpooling		
1×conv size: 3; ch: 384	3×conv size: 3; ch: 256	4×conv size: 3; ch: 256
	Maxpooling	
1×conv size: 3; ch: 384	3×conv size: 3; ch: 512	4×conv size: 3; ch: 512
	Maxpooling	
1×conv size: 3; ch: 256	3×conv size: 3; ch: 512	4×conv size: 3; ch: 512
Maxpooling		
Fully connected *fc6*, 4,096 neurons		
Fully connected *fc7*, 4,096 neurons		
Fully connected, 1,000 neurons		
Output: soft-max of probabilities for 1,000 object classes		

VGG16/VGG19

Whilst in AlexNet, the filter sizes change across the layers, in VGG16 and VGG19, all convolutional layers have a constant 3×3-sized receptive field [3]. Both deep architectures consist of 2 more max-pooling layers in comparison with AlexNet, and have deeper fully connected layers in cascade. Similar to AlexNet, the VGG architectures employ ReLUs for response normalisation. For the experiments described below, we use the 4,096 activations of the fully connected layers as feature vectors.

GoogLeNet

In contrast to AlexNet and VGG networks, GoogleNet uses so-called inception modules in succession (cf. Fig. 5.6). This module consists of parallel convolution layers and a max-pooling layer. The outputs of all layers are concatenated to produce a single output. The inception module thus collects multi-level features from every input on different scales. The activations of the last pooling layer are used as deep feature vectors.

5.2.2 Deep Convolutional Generative Adversarial Networks

Deep Convolutional Generative Adversarial Networks (DCGANs) are a type of Generative Adversarial Networks (GANs) which use CNNs in the generator and discriminator (cf. Sect. 5.1.2). Both the generator and discriminator consist of convolutional layers, but unlike CNNs, they do not have any pooling layers. DCGANs have been first introduced by Radford et al. [50] and have been applied for image classification tasks with state-of-the-art results [50]. DCGANs can be also used for unsupervised representation learning from audio data by training them on visual representations of the input audio signal, such as spectrograms or mel-spectrograms [51, 52].

5.2.2.1 System Structure

In this chapter, we introduce a DCGAN used in [51] for audio recognition tasks which has a less complex architecture than the DCGAN proposed by Radford et al. [50], who use $N_{layer}^{DCGAN} = 4$ and $N_{maps}^{DCGAN} = 64$. N_{layer}^{DCGAN} denotes the number of convolutional layers in the generator and discriminator CNNs, and N_{maps}^{DCGAN} is the number of the feature maps in the output layer of the generator. The structure used in [51] has 2 convolutional layers and 32 feature maps in the output layer of the generator. This simpler structure for audio analysis has been applied for three main reasons. First, the data employed for the audio task is much less than the data used in Radford et al.'s experiments [50]. Second, for the decreased amount of training data, the number of free parameters is reduced. Third, in [51], grayscale spectrograms with one channel have been used, whilst Radford et al. train their DCGANs on colour images with three channels.

For the introduced DCGAN in this work, both the generator and discriminator comprise the equal number N_{layer}^{DCGAN} of convolutional layers with a fixed stride of two. The output layer of the generator and the input layer of the discriminator have the spatial dimensions of the input (mel-)spectrograms that should be processed and contain N_{maps}^{DCGAN} feature maps. Figure 5.9 shows the structure of the DCNN generator. In each layer directly below the layer in the generator or on top of the layer in the discriminator, the number of the feature maps is doubled and the spatial dimensions are halved. In the final step, a 100-dimensional uniform noise distribution is applied as input to the generator, where it is projected to the dimensionality required by the first convolutional layer. As suggested by Radford et al., the feature maps have kernels with size 5×5 [50]. The structure of the DCNN generator in the DCGAN used in this chapter is given in Fig. 5.9. We analyse the efficacy of the representation learnt by the introduced DCGAN for three audio-based medical applications in Sect. 5.3.

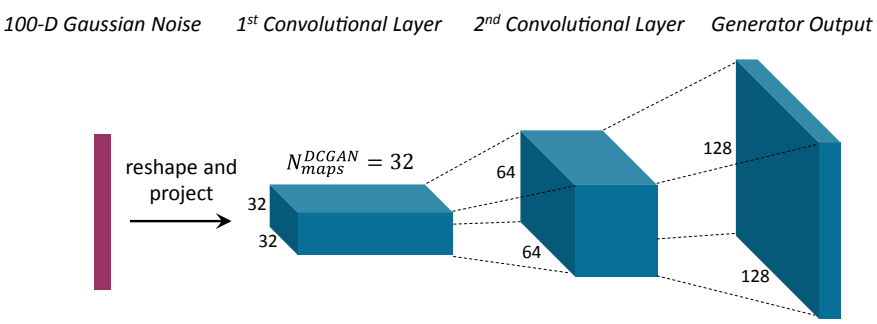


Fig. 5.9 Illustration of the generator in a DCGAN with $N_{layer}^{DCGAN} = 2$ and $N_{maps}^{DCGAN} = 32$ applied for generating spectrograms with hypothetical dimensions 128×128. A 100 dimensional Gaussian noise is projected and reshaped to a spatial convolutional representation. In every convolutional layer below the output layer, the spatial representations are halved. The convolutional layer directly below the output layer comprises N_{maps}^{DCGAN} feature maps. The number of the feature maps is then doubled in each further layer. The discriminator mirrors the DCNN architecture of the generator, and information flow would be right-to-left in the illustration. The number of convolutional layers in both generator and discriminator is similar

5.2.3 Recurrent Sequence to Sequence Autoencoders

Acoustic sequences are typically varying length signals; this highlights a drawback for CNNs-based deep representation learning architectures which generally require inputs of fixed dimensionality. Moreover, many DNN systems applied for representation learning, including Restricted Boltzmann Machines (RBMs) or stacked autoencoders, do not explicitly account for the inherent sequential nature of audio signals [53]. To learn fixed-length representations from variable-length data with sequential nature, sequence to sequence learning with recurrent neural networks (RNNs) has been proposed in machine translation [54–56]. Sequence to sequence autoencoders (S2SAEs) have been used for unsupervised pre-training of RNNs with state-of-the-art results on image recognition or text classification tasks [57]. Variational S2SAEs have been employed to learn representations of sentences, and to create new sentences from the latent space [58, 59]. Furthermore, Weninger et al. used denoising recurrent autoencoders to learn variable-length representations of audio for reverberated speech recognition [60].

5.2.3.1 AUDEEP—Autoencoder Structure

In this chapter, we introduce AUDEEP,[3] a recurrent S2SAE which can learn deep representations of time series data by taking into account their temporal dynamics [61, 62]. The advantages of this system have been shown in tasks like environmental

[3] https://github.com/auDeep/auDeep.

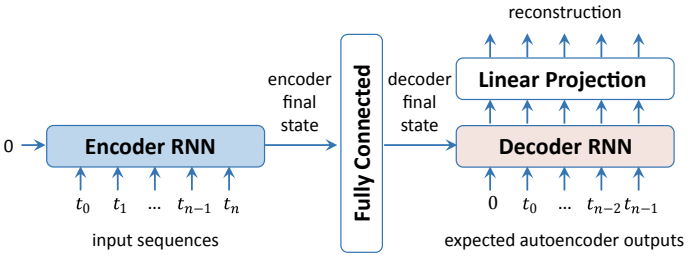


Fig. 5.10 A high-level overview of the recurrent autoencoder used in AUDEEP

and acoustic sound classification [61, 62], and abnormal heartbeat recognition [63]. The implementation of AUDEEP extends the RNN encoder-decoder model proposed by Sutskever et al. [54]. The high-level structure of the autoencoder used in AUDEEP is shown in Fig. 5.10. The input sequence is sent to a multilayered encoder RNN which collects important information about the input sequence in its hidden states. The final hidden state of the encoder RNN is then forwarded across a fully connected layer. The output of this fully connected layer is then used for initialising the hidden state of the multilayered decoder RNN. The decoder RNN has the task of reconstructing the input sequence based on the information contained in the initial hidden state. In the training phase, the network tries to minimise the root mean squared error (RMSE) between the input sequence and its reconstruction. After the autoencoder training is finished, the activations of the fully connected layer are used as the deep representation of the input sequence.

5.2.3.2 AUDEEP—Practical Usage

A high-level overview of AUDEEP is given in Fig. 5.11. First, spectrograms are generated from audio signals (cf. Fig. 5.11a). A sequence to sequence autoencoder, as described in Sect. 5.2.3.1, is then trained on the generated spectrograms (cf. Fig. 5.11b). The process of spectrogram creation is described in Sect. 5.2.1.2. Afterwards, the learnt representation of each input instance is extracted as its feature vector (cf. Fig. 5.11c). Finally, if instance labels are available, a classifier can then be

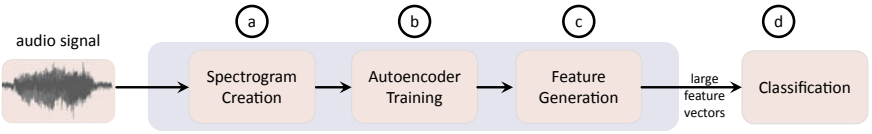


Fig. 5.11 Structure of AUDEEP for deep representation learning using sequence to sequence autoencoders. The autoencoder training is entirely unsupervised

trained and evaluated on the obtained features (cf. Fig. 5.11d). In Sect. 5.3, we analyse the efficacy and robustness of AUDEEP representations for medical applications.

5.3 Medical Applications

The capabilities of the proposed unsupervised feature learning techniques are exemplified in three different audio recognition tasks from the field of health care: recognition of heart sound abnormalities, classification of snore sounds, and recognition of bipolar disorder. For each task, we first describe its dedicated public corpus. Subsequently, in the experimental settings sections, the configurations and hyperparameters applied for each task are defined. Finally, the results are shown and analysed. The evaluation metric for all tasks is Unweighted Average Recall (UAR). We use UAR, as this measure gives equal weight to all classes of each corpus and is accordingly more suitable than a weighted metric (e.g., accuracy) for databases which have imbalanced class ratio.

5.3.1 Abnormal Heartbeat Recognition

For the first experimental evaluation of the discussed methods, the **Heart Sounds Shenzhen (HSS)** corpus was used. In the following, the data collection process and the partitioning are described, the experimental settings of all feature learning approaches and two baseline methods, using established acoustic feature sets, are introduced, and the results are given.

5.3.1.1 Heart Sounds Shenzhen Corpus

The HSS corpus has been published through the Interspeech 2018 Computational Paralinguistics ChallengE (ComParE) [64], an annual scientific machine learning competition, organised as a special session of the Interspeech conference. HSS contains 845 audio recordings of approximately 30 s each, totalling up to 422.8 min. The corpus was collected from 170 subjects (55 female, 115 male), aged from 21 to 88, an average age of 65.4 years, with a standard deviation of 13.2 years, i.e., mostly elderly individuals. Health conditions vary across subjects, including coronary heart disease, heart failure, arrhythmia, hypertension, hyperthyroid, and valvular heart disease. The data was recorded in four different locations using an electronic stethoscope (sampling rate 4 kHz): (i) the auscultatory mitral area, (ii) the pulmonary valve auscultation area, (iii) the pulmonary valve auscultation area, and (iv) the auscultatory area of the triscuspid valve. All instances have been categorised by specialists into the following three classes: (1) *normal* heart beat, (2) *mild* abnormality, or (3) *severe* abnormality. The diagnosis was verified by an echocardiography (cardiac

Table 5.2 Heart sound dataset. Class distribution per partition

Class	Training	Development	Test	Sum
Normal	84	32	28	144
Mild	276	98	91	465
Moderate/severe	142	50	44	236
Sum	502	180	163	845

ultrasound). The recordings were divided into subject-disjunct training, development, and test partitions, resulting in the class distribution shown in Table 5.2.

5.3.1.2 Experimental Settings

In the experiments presented below, two *non-deep learning* baseline approaches are evaluated and compared to the introduced deep learning-based frameworks. Two types of well-established hand-crafted acoustic features are used in combination with a *support vector machine (SVM)* classifier. First, the openSMILE [65] ComParE acoustic feature set is employed [66], which is a large-scale brute-forced set, comprising 6,373 acoustic descriptors. The features are based on 130 *low-level descriptors (LLDs)*, such as *energy*, *pitch*, *MFCCs*, *spectral* descriptors and a certain number of functionals (statistics, such as *moments*, *percentiles*, and *extrema*) applied to the LLDs, in order to summarise them over the whole audio instance. Second, the 130 LLDs are represented as *bag-of-audio-words* (BoAW), i.e., a histogram representation of the LLDs occuring in a single chunk. For this, the LLD vectors extracted from short frames of 20–60 ms width are quantised according to a pretrained codebook; then, a histogram is generated counting the frequencies of each codebook 'audio word' in the corresponding audio instance. Finally, the logarithm is taken from the term frequencies to compress their range. Acoustics LLDs as well as functionals are extracted with the toolkit openSMILE [65], BoAW are generated with the tool openXBOW[4] [67]. For BoAW, the size of the codebook, generated through a sub-sampling of the LLDs from the training set, is a critical hyperparameter of the approach and is optimised using the development partition. As the class label is not taken into account in this process, BoAW is also an unsupervised feature learning method, though it highly depends on a suitable selection of hand-crafted LLDs. Moreover, the *complexity* parameter of the SVM is optimised on this partition, whilst the kernel is *linear*, as no improvement is found with other kernels for the considered acoustic features. More details on the experimental settings are found in the works by Amiriparian et al. [63] and Schuller et al. [64].

[4] https://github.com/openXBOW/openXBOW.

DEEP SPECTRUM

To obtain the DEEP SPECTRUM features, mel-spectrograms with the window width of 1 s and the window hop of 0.5 s are forwarded through all pre-trained networks. Afterwards, the activations of the neurons on the second fully connected layer (*fc7*) of AlexNet, VGG16/VGG19, and the activations of the last pooling layer of GoogLeNet are extracted as feature vectors. Finally, a linear SVM is applied to classify the corpus.

DCGAN

The DCGAN architecture is chosen based on the results reported by Radford et al. [50], who use $N_{layer}^{DCGAN} = 4$ and $N_{maps}^{DCGAN} = 64$. A less complex DCGAN architecture with $N_{layer}^{DCGAN} = 3$ and $N_{maps}^{DCGAN} = 32$ is applied here. The DCGAN is trained on mel-spectrograms with the window width of 0.16 s, the window hop of 0.08 s, and $N_{mel} = 256$ mel frequency bands, with amplitude clipping. We used amplitude thresholds below $\{-30, -45, -60, -70\}$ dB. From each of the four configurations and their fusion, a feature vector has been extracted and evaluated.

AUDEEP

The S2SAEs are trained on mel-spectrograms extracted with the window width of 0.32 s, the window overlap 0.16 s, and $N_{mel} = 128$ mel frequency bands, with amplitude threshold under $\{-30, -45, -60, -70\}$ dB. For the autoencoder $N_{layer} = 2$ layers and $N_{unit} = 256$ Gated Recurrent Units (GRUs) with a unidirectional encoder RNN and a bidirectional decoder RNN are applied.

5.3.1.3 Results and Discussion

The detailed classification results obtained from the deep learning approaches are shown in Table 5.3. We also compare these results with two non-deep learning baseline systems, COMPARE and BOAW, used in [63, 64]. In Table 5.4, results on the development and test partitions of all five approaches are shown for the configuration performing best on the development partition.

The HSS corpus opposes a big machine learning challenge because of its relatively short length and high acoustic similarities between the three classes. However, the deep learning approaches were able to learn strong representations and match or improve upon the results obtained from the expert-designed feature sets. All three approaches show better results on the test set than the development partition, whilst the the baseline system COMPARE seems to be overfitted on the training data. Such an overfitted model will be less precise on unseen data than a more generally fitted model.

Table 5.3 Classification results for the HSS corpus. UAR of the three classes (normal, mild, and moderate/severe) is used as scoring metric. Best result is highlighted in bold with a grey shading. The chance level is 0.333 UAR. The power levels (dB) specify the threshold, below which the signals are clipped

DEEP SPECTRUM					DCGAN					AUDEEP				
System	Dim.	C	Devel.	Test	System	Dim.	C	Devel.	Test	System	Dim.	C	Devel.	Test
AlexNet	4,096	10^{-2}	0.441	0.461	−30 dB	2,048	10^{-3}	0.340	0.411	−30 dB	1,024	$2 \cdot 10^{-2}$	0.328	0.400
VGG16	4,096	10^{-2}	0.418	0.429	−45 dB	2,048	10^{-3}	0.397	0.425	−45 dB	1,024	$5 \cdot 10^{-4}$	0.384	0.406
VGG19	4,096	10^{-1}	0.427	0.459	−60 dB	2,048	10^{-3}	0.400	0.462	−60 dB	1,024	$6 \cdot 10^{-2}$	0.396	0.452
GoogLeNet	1,024	10^{-1}	0.403	0.411	−75 dB	2,048	10^{-3}	0.402	0.443	−75 dB	1,024	$8 \cdot 10^{-3}$	0.369	0.417
–	–	–	–	–	Fused	8,192	10^{-3}	0.412	0.460	Fused	4,096	$4 \cdot 10^{-3}$	0.352	**0.479**

Table 5.4 Comparison of the best performing configuration of each approach on the development partition of the HSS corpus. Best result is high-lighted in bold with a grey shading

UAR	Deep Spectrum	DCGAN	AUDEEP	COMPARE	BOAW
Development	0.441	0.412	0.396	0.503	0.437
Test	0.461	0.460	0.452	**0.464**	0.410

Table 5.5 Munich-Passau snore sound corpus. Class distribution per partition

Class	Training	Development	Test	Sum
V	168	161	155	484
O	76	75	65	216
T	8	15	16	39
E	30	32	27	89
Sum	282	283	263	828

5.3.2 Snore Sound Classification

As a second example, experiments are conducted on a corpus of snore sound recordings. Automatic classification of snore types from audio is relevant for medical diagnostics, as some types require medical treatment. The usual procedure is to undertake a drug-induced sleep endoscopy. As opposed to that, a classification based on a sound recording during natural sleep would be completely stress-free for the patient. As in the preceding Section, the data, experimental settings, and results are described in the following paragraphs.

5.3.2.1 Munich-Passau Snore Sounds Corpus

The **Munich-Passau Snore Sounds Corpus (MPSSC)** [49] consists of 828 snore events from 219 subjects, split into subject-disjunct training, development, and test partitions. Each snore event has been classified into one out of four snore types, depending on the location of obstruction in the throat: *V (velum)*, *O (oropharynx)*, *T (tongue)*, *E (epiglottis)*. The type of snoring is usually constant across subjects. Table 5.5 shows the number of snore events (instances) per class and partition. The duration of single snore events ranges from 0.3 to 2.0 s. The corpus was used for benchmarking in ComParE 2017 [68].

5.3.2.2 Experimental Settings

The DCGAN and AUDEEP approaches and the COMPARE and BOAW baseline systems are trained and evaluated with the same configurations as introduced

in Sect. 5.3.1.2. For extraction of the DEEP SPECTRUM features, similar configurations as described in [43] are applied, i.e., spectrograms from a complete recording without sliding windows are fed into the pre-trained CNNs.

5.3.2.3 Results and Discussion

In Table 5.6, the results of the deep learning experiments on the MPSSC are presented. These approaches are also compared with the COMPARE and BOAW baselines. Results on the development and test partitions of all five approaches for the configuration performing best on the development set are shown in Table 5.7. AlexNet features achieved the best results on both training and test partitions of the MPSSC. The classification results using DCGAN features, except for the threshold −75 dB, are comparable with the baseline results. We also observe that there is a small mismatch between the training and test partitions of the dataset (cf. DCGAN −45 dB and DCGAN −75 dB, or AUDEEP −45 dB and AUDEEP −60 dB). The DEEP SPECTRUM features obtained from VGG16, VGG19, and GoogLeNet show relatively weak performance on the development partition, but achieve comparable results with the baseline's test set. All AUDEEP features show stronger performance than test results of the baseline systems.

5.3.3 Bipolar Disorder Recognition

Finally, the feature learning methods are evaluated on a speech database of subjects with different levels of bipolar disorder. The remainder of this section follows the structure of the preceding ones.

5.3.3.1 Bipolar Disorder Corpus

Patients suffering from *bipolar disorder (BP)* are usually experiencing intermittent episodes of mania and depression. With each phase lasting from days to months, subjects show changing levels of mood and activity, and are affected in their ability to carry out tasks. From the Turkish audio-visual **Bipolar Disorder Corpus (BDS)** [69, 70], the audio recordings were used for the experiments. The data consists of structured interviews of 46 Turkish subjects (15 female/30 male, age ranging from 18 to 60), all suffering from BP and experiencing the mania episode. Patients had to complete seven tasks, including telling a happy and a sad memory and explaining emotion eliciting pictures.

Each recording session was labelled according to the level of mania shown (*remission*, *hypomania*, or *mania*). The data was split into three partitions (training, development, and test), balancing age and gender. Details on the distribution of subjects and overall duration for each partition and class are shown in Table 5.8.

Table 5.6 Classification results for the MPSSC. UAR of the four classes (**v**elum, **o**ropharynx, **t**ongue, and **e**piglottis) is used as scoring metric. Best result is highlighted in bold with a grey shading. The chance level is 0.250 UAR. The power levels (dB) specify the threshold, below which the signals are clipped

DEEP SPECTRUM					DCGAN					AUDEEP				
System	Dim.	C	Devel.	Test	System	Dim.	C	Devel.	Test	System	Dim.	C	Devel.	Test
AlexNet	4,096	10^{-4}	0.448	**0.670**	−30 dB	1,536	10^{-2}	0.386	0.540	−30 dB	1,024	10^{-1}	0.443	0.617
VGG16	4,096	10^{-4}	0.315	0.541	−45 dB	1,536	10^{-3}	0.329	0.531	−45 dB	1,024	10^{-1}	0.425	0.591
VGG19	4,096	10^{-3}	0.326	0.525	−60 dB	1,536	10^{-2}	0.371	0.561	−60 dB	1,024	10^{-1}	0.434	0.569
GoogLeNet	1,024	10^{-1}	0.285	0.510	−75 dB	1,536	10^{-3}	0.351	0.465	−75 dB	1,024	10^{-1}	0.374	0.596
–	–	–	–	–	Fused	6144	10^{-1}	0.351	0.579	Fused	4,096	10^{-1}	0.448	0.613

Table 5.7 Comparison of the best performing configuration of each approach on the development partition of the MPSSC. Best result is high-lighted in bold with a grey shading

UAR	DEEP SPECTRUM	DCGAN	AUDEEP	COMPARE	BOAW
Development	0.448	0.386	0.448	0.406	0.466
Test	**0.670**	0.540	0.613	0.585	0.499

Table 5.8 Bipolar Disorder Corpus. Number and total duration (minutes:seconds) of recordings per class and partition. Details on the test partition are blinded as, at the time of writing, the data was an active challenge dataset

Class	Training	Development	Test
Remission	25–64:52	18–42:47	–
Hypomania	38–167:42	21–62:24	–
Mania	41–189:29	21–71:01	–
Sum	104–422:04	60–176:13	54–207:07

5.3.3.2 Experimental Settings

For this task, we decided to run stratified four-fold cross-validation for classification instead of partitioning, because of the class distribution and the number of subjects. Due to this, we cannot directly compare our results with the baselines reported in [70].

DEEP SPECTRUM

We have applied the same configurations as described in Sect. 5.3.1.2 for extraction of the DEEP SPECTRUM features.

DCGAN

The DCGAN is trained on mel-spectrograms with the window width of 1 s, the window hop of 0.5 s, and $N_{mel} = 256$ mel frequency bands. The DCGAN configurations remain unchanged (cf. Sect. 5.3.1.2).

AUDEEP

The S2SAEs are trained on mel-spectrograms extracted with the window width of 0.8 s, the window overlap 0.4 s, and $N_{mel} = 256$ mel frequency bands. The autoencoder settings remain unchanged (cf. Sect. 5.3.1.2).

Table 5.9 Classification results for the AVEC 2018 BDS. UAR of the three classes of BD (remission, hypo-mania, and mania) is used as scoring metric. A stratified four-fold cross-validation have been applied for classification. Best result is highlighted in bold with a grey shading. The chance level is 0.333 UAR

DEEP SPECTRUM				DCGAN				AUDEEP			
System	Dim.	C	4-fold CV	System	Dim.	*C*	4-fold CV	System	Dim.	*C*	4-fold CV
AlexNet	4,096	10^{-3}	0.366 (± 0.015)	−30 dB	2,048	10^{-2}	0.414 (± 0.150)	−30 dB	1,024	10^{-2}	0.468 (± 0.111)
VGG16	4,096	10^{-3}	0.359 (± 0.019)	−45 dB	2,048	10^{-2}	0.424 (± 0.179)	−45 dB	1 024	10^{-2}	0.498 (± 0.221)
VGG19	4,096	10^{-3}	0.360 (± 0.022)	−60 dB	2,048	10^{-2}	**0.498 (± 0.089)**	−60 dB	1,024	10^{-2}	0.479 (± 0.071)
GoogLeNet	1,024	10^{-3}	0.349 (± 0.032)	−75 dB	2,048	10^{-2}	0.433 (± 0.124)	−75 dB	1,024	10^{-2}	0.427 (± 0.098)
–	–	–	–	Fused	8,192	10^{-2}	0.432 (± 0.188)	Fused	4,096	10^{-2}	0.459 (± 0.083)

5.3.3.3 Results and Discussion

Detailed results of the deep learning experiments on the AVEC 2018 BDS are given in Table 5.9. The provided UAR is the average of the UARs of all folds. The best classification result is achieved with features learnt by the DCGAN system with the threshold of −60 dB for amplitude clipping. Both AUDEEP and DCGAN show similar performance on this task, whilst the DEEP SPECTRUM results lay behind them. We assume that the mel-spectrograms applied for the DEEP SPECTRUM experiments do not contain enough discriminative information for each of the three bipolar classes.

5.4 Conclusions

The human body produces a wide range of acoustic sounds that directly and indirectly reflect developments and changes in our pathological and physiological states. These acoustic data are complex in nature, exhibit strong interconnections, and are rare. For automatic analysis of such data, sophisticated machine learning systems are required which can cope with the mentioned restrictions and challenges. Conventional machine leaning methods need precise engineering and fundamental domain knowledge to design features from which a machine learner can identify the patterns of the input data. This manual process of feature engineering can be tedious and costly due to the large amount of human intervention. Whilst task-specific features frequently outperform more general feature sets [71, 72], they may not show strong performance for unrelated tasks. On the contrary, representation learning approaches are able to learn task-dependent, robust features directly from the raw data [51, 53, 63, 73–75].

In order to cope with the data scarcity and mentioned machine learning challenges, especially for medical applications, we have introduced three state-of-the-art deep learning approaches. In Sect. 5.1, we explained the theoretical background of the core neural network architectures applied in our deep learning systems. We have characterised the fundamental structures of CNNs, RNNs, autoencoders, and GANs. In Sect. 5.2, we have introduced DEEP SPECTRUM (cf. Sect. 5.2.1), a DCGAN (cf. Sect. 5.2.2), and AUDEEP (cf. Sect. 5.2.3), and described their technical structures. We then exemplified and evaluated the capabilities of the proposed techniques in three different audio recognition tasks from the field of health care: recognition of heart sound abnormalities (cf. Sect. 5.3.1), classification of snore sounds (cf. Sect. 5.3.2), and recognition of bipolar disorder (cf. Sect. 5.3.3).

We showed that using the proposed CNN-, and RNN-based methodologies, it is possible to improve upon conventional machine learning techniques which use brute-force, expert-designed features. Whilst CNNs have been widely applied in audio processing [74–77], deep pre-trained CNNs make a compelling case to be considered as a legitimate and stable audio feature extraction technique. Especially for analysing medical data which is scarce, transfer learning utilising pre-trained models shows strong promise (cf. Sect. 5.3). We also demonstrated that adversar-

ial networks can learn robust representations from spectral features. The introduced DCGAN is also capable of generating spectrograms which are highly similar to training data. This finding can help in extending the DCGAN framework for synthesising acoustic medical data.

Unlike DEEP SPECTRUM and DCGAN approaches which require inputs of fixed dimensionality, AUDEEP is able to learn a fixed-length representation from variable length acoustic signals, whilst considering their time-dependent nature. As all three systems are unsupervised feature extraction and learning techniques, they are more suitable for data of heterogeneous nature. They fit well for (real-time) big data analysis, and are less susceptible to overfitting.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Res. Repos. (CoRR) (2014), arXiv:abs/1409.1556
4. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725–1732 (2014)
5. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1510–1517 (2018)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE (2017)
7. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. BMVC **1**(3), 6 (2015)
8. Farfade, S.S., Saberian, M.J., Li, L.-J.: Multi-view face detection using deep convolutional neural networks. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 643–650. ACM (2015)
9. Li, H., Lin, Z., Shen, X., Brandt J., Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)
10. Boureau, Y.-L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 111–118 (2010)
11. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems, pp. 649–657 (2015)
12. Kim, Y.: Convolutional neural networks for sentence classification (2014). arXiv:1408.5882
13. Poria, S., Peng, H., Hussain, A., Howard, N., Cambria, E.: Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing **261**, 217–230 (2017)
14. Wehrmann, J., Becker, W., Cagnini, H.E., Barros, R.C.: A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In: Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2384–2391. IEEE (2017)

15. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017)
16. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE (2015)
17. Valenti, M., Diment, A., Parascandolo, G., Squartini, S., Virtanen, T.: Dcase 2016 acoustic scene classification using convolutional neural networks. In: IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016). Budapest, Hungary (2016)
18. Lidy, T., Schindler, A.: Cqt-based convolutional neural networks for audio scene classification. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), vol. 90. DCASE2016 Challenge, pp. 1032–1048 (2016)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
20. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2172–2180 (2016)
21. LeCun, Y.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
22. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017). arXiv:1701.07875
23. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks (2017). arXiv:1701.04862
24. Valle, R., Cai, W., Doshi, A.: Tequilagan: how to easily identify gan samples (2018). arXiv:1807.04919
25. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
26. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318 (2013)
27. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
28. Hochreiter, S., Schmidhuber, J.: Lstm can solve hard long time lag problems. In: Advances in Neural Information Processing Systems, pp. 473–479 (1997)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
30. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine learning, pp. 1096–1103. ACM (2008)
31. Lee, C.H., Yoon, H.-J.: Medical big data: promise and challenges. Kidney Res. Clin. Pract. **36**(1), 3 (2017)
32. Topol, E.J.: The big medical data miss: challenges in establishing an open medical resource. Nat. Rev. Genet. **16**(5), 253 (2015)
33. Vithanwattana, N., Mapp, G., George, C.: mhealth-investigating an information security framework for mhealth data: challenges and possible solutions. In: Proceedings of the 12th International Conference on Intelligent Environments (IE), pp. 258–261. IEEE (2016)
34. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE (2009)
35. Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S.: Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Comput. **22**(2), 511–538 (2010)
36. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 609–616. ACM (2009)
37. Pinto, N., Doukhan, D., DiCarlo, J.J., Cox, D.D.: A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput. Biol. **5**(11), e1000579 (2009)

38. Jarrett, K., Kavukcuoglu, K., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: Proceedings of the 12th IEEE International Conference on Computer Vision, pp. 2146–2153. IEEE (2009)
39. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
40. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: ISBI, pp. 294–297. Citeseer (2015)
41. Hoo-Chang, S., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285 (2016)
42. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
43. Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Schuller, B.: Snore sound classification using image-based deep spectrum features. In: Proceedings of Interspeech, pp. 3512–3516. Stockholm, Sweden (2017)
44. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE, Boston, MA, USA (2015)
45. Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Pugachevskiy, S., Schuller, B.: Bag-of-deep-features: noise-robust deep feature representations for audio analysis. In: Proceedings of the 31st International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 2419–2425. IEEE, Rio de Janeiro, Brazil (2018)
46. Amiriparian, S., Cummins, N., Gerczuk, M., Pugachevskiy, S., Ottl, S., Schuller, B.: Are you playing a shooter again?! Deep representation learning for audio-based video game genre recognition. IEEE Trans. Games **11**, 11 pages (2019), to appear
47. Amiriparian, S., Cummins, N., Ottl, S., Gerczuk, M., Schuller, B.: Sentiment analysis using image-based deep spectrum features. In: Proceddings of the 2nd International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2017) held in conjunction with the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017), AAAC, pp. 26–29. IEEE, San Antonio, TX (2017)
48. Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., Schuller, B.: An image-based deep spectrum feature representation for the recognition of emotional speech. In: Proceedings of the 25th ACM International Conference on Multimedia, MM 2017, ACM. ACM, Mountain View, CA (2017)
49. Janott, C., Schmitt, M., Zhang, Y., Qian, K., Pandit, V., Zhang, Z., Heiser, C., Hohenhorst, W., Herzog, M., Hemmert, W., Schuller, B.: Snoring classified: the munich-passau snore sound corpus. Comput. Biol. Med. **94**, 106–118 (2018)
50. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015). arXiv:1511.06434
51. Amiriparian, S., Freitag, M., Cummins, N., Gerzcuk, M., Pugachevskiy, S., Schuller, B.W.: A fusion of deep convolutional generative adversarial networks and sequence to sequence autoencoders for acoustic scene classification. In: Proceedings of the 26th European Signal Processing Conference (EUSIPCO), EURASIP, 5 pages. IEEE, Rome, Italy (2012), to appear
52. Chang, J., Scherer, S.: Learning representations of emotional speech with deep convolutional generative adversarial networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2746–2750. IEEE (2017)
53. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
54. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, Curran Associates, Inc., pp. 3104–3112 (2014)

55. Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. ACL, Doha, Qatar (2014)
56. Luong, M.-T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning, 10 pages (2015). arXiv:1511.06114
57. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pp. 3079–3087 (2015)
58. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space, 12 pages (2015). arXiv:1511.06349
59. Jang, M., Seo, S., Kang, P.: Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning (2018). arXiv:1802.03238
60. Weninger, F., Watanabe, S., Tachioka, Y., Schuller, B.: Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4623–4627. IEEE (2014)
61. Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., Schuller, B.: audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. J. Mach. Learn. Res. **18**(1), 6340–6344 (2017)
62. Amiriparian, S., Freitag, M., Cummins, N., Schuller, B.: Sequence to sequence autoencoders for unsupervised representation learning from audio. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), IEEE, pp. 17–21. IEEE, Munich, Germany (2017)
63. Amiriparian, S., Schmitt, M., Cummins, N., Qian, K., Dong, F., Schuller, B.: Deep unsupervised representation learning for abnormal heart sound classification. In: Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2018, IEEE, 4 pages. IEEE, Honolulu, HI (2018) to appear
64. Schuller, B., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F.B., Rathner, E.-M., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., Zafeiriou, S.: The interspeech 2018 computational paralinguistics challenge: addressee, cold & snoring. In: Proceedings of Interspeech, pp. 122–126. ISCA, Hyderabad, India (2018)
65. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of ACM Multimedia, pp. 835–838. ACM, Barcelona, Catalunya, Spain (2013)
66. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of Interspeech, pp. 148–152. ISCA, Lyon, France (2013)
67. Schmitt, M., Schuller, B.: Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit. J. Mach. Learn. Res. **18**(1), 1–5 (2017)
68. Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A.S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., Tzirakis, P., Zafeiriou, S.: The interspeech 2017 computational paralinguistics challenge: atypical & self-assessed affect, crying & heart beats. In: Proceedings of Interspeech, pp. 3442–3446. ISCA, Stockholm, Sweden (2017)
69. Çiftçi, E., Kaya, H., Güleç, H., Salah, A.A.: The turkish audio-visual bipolar disorder corpus. In: Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). AAAC, Beijing, China (2018)

70. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Çiftçi, E., Güleç, H., Salah, A.A., Pantic, M.: Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In: Proceedings of the 8th Annual Workshop on Audio/Visual Emotion Challenge. ACM, Seoul, Korea (2018), to appear
71. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andrè, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
72. Huang, Z., Dang, T., Cummins, N., Stasak, B., Le, P., Sethu, V., Epps, J.: An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 41–48. ACM, Brisbane, AU (2015)
73. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press Cambridge (2016)
74. Amiriparian, S., Julka, S., Cummins, N., Schuller, B.: Deep convolutional recurrent neural networks for rare sound event detection. In: Proceedings 44. Jahrestagung für Akustik, DAGA 2008, DEGA, pp. 1522–1525. DEGA, Munich, Germany (2018)
75. Amiriparian, S., Baird, A., Julka, S., Alcorn, A., Ottl, S., Petrović, S., Ainger, E., Cummins, N., Schuller, B.: Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH 2018, pp. 2334–2338. ISCA, Hyderabad, India (2018)
76. Bae, I., Choi, S.H., Kim, N.S.: Acoustic scene classification using parallel combination of LSTM and CNN. In: Proceedings of DCASE'16, satellite to EUSIPCO'16, pp. 11–15. IEEE (2016)
77. Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B., Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings of ICASSP'16, pp. 5200–5204. IEEE, Shanghai, P. R. China (2016)

Part II
Augmentation

Chapter 6
Data Augmentation in Training Deep Learning Models for Medical Image Analysis

Behnaz Abdollahi, Naofumi Tomita and Saeed Hassanpour

Abstract Data augmentation is widely utilized to achieve more generalizable and accurate deep learning models based on relatively small labeled datasets. Data augmentation techniques are particularly critical in medical applications, where access to labeled data samples is commonly limited. Although data augmentation methods generally have a positive impact on the performance of deep learning models, not all data augmentation techniques are applicable and suitable for analyzing medical images. In this chapter, we review common image augmentation techniques and their properties. Furthermore, we present and evaluate application-specific data augmentation methods that are beneficial for medical image analysis. The material presented in this chapter aims to guide the use of data augmentation techniques in training deep learning models for various medical image analysis applications, in which annotated data are not abundant or are difficult to acquire.

6.1 Introduction

Deep learning methods have shown the state-of-the-art results for image classification and object detection, and in some cases, have even exceeded human performance [1]. These deep learning methods rely on a network of multiple computational layers to model high-level abstractions in the data [2]. The high-level abstract representations are instrumental for data analysis tasks, such as image classification and object detection. Deep learning is strongly rooted in previously existing artificial neural networks [3]; however, the construction of deep learning models only became practical in the last decade due to the availability of large training data sources, such as ImageNet with millions of annotated images [4].

With the recent expansions in radiological image repositories and digitized histology archives, the field of medical image analysis is ripe for the development and application of deep learning models to assist clinicians in performing diagnosis and

B. Abdollahi · N. Tomita · S. Hassanpour (✉)
Biomedical Data Science Department, Dartmouth College, Hanover, NH 03755, USA
e-mail: Saeed.Hassanpour@dartmouth.edu

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_6

prognosis, and with managing patients. However, a major bottleneck in the development and use of deep learning methods in medical applications is the lack of large annotated datasets that are required for training deep learning models. A substantial amount of time and medical expertise is required for data annotation, which hinders such data collection efforts. Of note, in contrast to general image classification and object detection, generating labeled data for medical image analysis requires the consensus of opinion of multiple clinical domain-experts, and that process is time-consuming and resource intensive. In addition, concerns regarding patient privacy and data security have contributed to the scarcity of large, labeled medical image datasets in this field.

For training deep learning models on relatively small labeled datasets, data augmentation techniques have been an instrumental data-processing step to achieve a more generalizable and accurate performance. These data augmentation methods generate new annotated data points for training, based on existing data, through transformations that add small variations to the original data points while preserving their labels. Also, introducing additional variability in training sets through data augmentation has a regularization effect on the training process and guards the resulting models against overfitting.

Model robustness is another essential characteristic of an effective machine learning model. A deep learning model is considered robust if the testing error is consistent with the training error, and the model performance is stable in the presence of noise in the dataset. In practice, robustness is a critical requirement for training models on inconstant and noisy datasets such as medical images. Leveraging data augmentation methods can improve model robustness, as well as accuracy.

Although data augmentation methods are typically considered to have a positive impact on deep learning models and increase the overall accuracy of their results, not all data augmentation techniques are appropriate and helpful in medical applications. In this chapter, we review common image augmentation techniques and their application-specific properties. In addition, we present and evaluate various data augmentation methods, such as random rotation, random translation, random inversion along horizontal/vertical axes, elastic distortion [5], and color jittering [6], in detail, for various medical image analysis applications.

6.2 Data Augmentation Methodology

Data augmentation methods entail a set of techniques that artificially expand the size of a dataset for training machine learning models. Data augmentation has earned major attention in deep learning applications, especially after the advent of deep convolutional neural networks (CNNs), which consist of multiple convolutional layers [6–8]. The CNN architectures typically have significantly more model parameters than data samples that are available for training. Thus, deep learning models without proper regularization are prone to overfit the training data. To alleviate overfitting and achieve better prediction performance on unseen data samples (i.e., generalizability),

geometric transformations are applied on training data to augment and expand the dataset. Furthermore, recent studies have shown the effectiveness of interpolation-based techniques and generative models for data augmentation [9–11]. In this section, we review common data augmentation techniques that were originally developed for generic computer vision applications, in addition to some more recently published papers that are focused on analyzing medical images based on deep learning models.

6.2.1 Basic Augmentation

Many image augmentation techniques have been developed to make machine learning models robust by applying different transformations to input images. In this section, we review image cropping, flipping, affine transformations, and color perturbation as common transformation methods that are used for basic data augmentation.

One of the most basic and common data augmentation techniques is image cropping. Intuitively, a human observer can still recognize an object or a texture in magnified images, while from a neural network perspective, the input values are significantly different when their scale is changed. Therefore, as a result of such change in input values, the activation values in the inner layers of the network change as well. Image cropping is aimed to add scale invariance and robustness to the corpus of input images. The size of cropped images varies in different applications and depends on the resolution of original images, but the common practice is to randomly extract a specific window size from original images, followed by scaling it to a fixed size, which matches the input layer dimensions of the network. The position of the window is randomly selected at runtime.

Image flipping or mirroring is one of the simplest and most effective data augmentation techniques that is widely adopted in training deep learning models. The mirroring involves both horizontal and vertical axes. This technique is effective in many applications because in most cases flipped images are considered valid samples and preserve their original class annotations; however, convolutional neural networks can consider them as entirely new input (e.g., skin cancer classification [10] or skeletal maturity prediction [12, 13]). In some other cases, where anatomical features are tightly connected to their position in the original image, such as anatomical features for unilateral organs, flipping is not applicable.

Affine transformations, which are also commonly used for data augmentation, include a set of geometric transformations such as rotation, translation, and sheering. For data augmentation, one or a combination of these transformations with randomly selected transformation parameters can be applied to the input image. We review each of these transformations in more detail.

Image rotation, which geometrically rotates images by a certain degree, is commonly utilized as a basic data augmentation method, unless the orientation of target objects or input features are specific to a certain orientation (e.g., canonical orientation). Even for images with canonical orientations, a slight rotation of less than

30° can improve the generalizability of a model. Commonly used parameters for the rotation degree are based on random selection from a set of fixed degrees, such as 30°, 45°, or 90°. Particularly, 90° (a right-angle) rotation is one of the most common choices because (1) it does not produce a rotation artifact (i.e., background area outside the rotated image) and (2) it is computationally simple and fast.

Image translation moves each pixel in an image by a certain amount in x and y directions. This translation is described mathematically as follows:

$$I'(x + t_{x'} y + t_y) = I(x, y)$$

where I' and I are translated and original images, respectively; and t_x and t_y are displacements along x and y axes.

To effectively utilize the translation technique, we need to answer a fundamental question: how can a CNN that inherently has translation invariance properly benefit from translation? The local parameter sharing and the use of pooling layers in CNN architectures enable a network to have shift-invariance, which is a major strength of CNNs and also counter-acts the translation-based data augmentation. However, in some cases, the samples generated by translation can still improve the performance of the network. For example, for a lesion classification task in microscopic images, a translation operation can move the lesion in a way so that only a part of the lesion is visible in the augmented image. In this case, the deep neural network classifier is expected to make a correct prediction by considering a partial view of the lesion. This incompleteness in training data points has a regularization effect and improves the robustness of the networks. Of note, this effect is similar to the dropout technique used in hidden layers.

Other augmentation techniques have focused on changing the color distribution of images [6]. Color jittering randomly changes the contrast, saturation, brightness, and hue of a training image. Unlike natural image classification where a PCA-based jittering is found to be effective, the primary motivation for applying color jittering in the medical domain is to simulate observations under various lighting conditions and variations caused by different imaging and scanning hardware [14, 15].

6.2.2 Interpolation-Based Augmentation

Elastic deformation, a common augmentation technique in this category, was originally developed to mimic the oscillation of handwriting. Elastic deformation has been applied to grayscale medical image datasets such as CT scans, MRI images, and mammography images [10, 16–19]. This technique creates a deformation effect on an image by generating a random displacement field, whose intensity is controlled by the elasticity coefficient σ, and a scaling factor α [5]. This deformation is described mathematically as follows:

$$I'(x + \Delta x(x, y), y + \Delta y(x, y)) = I(x, y)$$

$$\Delta x = G(\sigma) * (a \times Unif([-1, 1, n, m]))$$

$$\Delta y = G(\sigma) * (a \times Unif([-1, 1, n, m]))$$

where I' and I are deformed and original images, Δx and Δy are horizontal and vertical displacement grids of n $\times$ m whose values are generated by a Gaussian filter with standard deviation of σ, applied on the product of uniformly drawn value in $[-1, 1]$ and α.

This technique is suitable for images capturing elastic objects, such as organs, skins, or soft tissues, which can naturally deform by a patient's breathing, movement, or the contact pressure of imaging hardware. Although the ranges of σ and α parameters are mostly arbitrary, a grid search along with visual validation on training images is useful to determine proper ranges for these parameters for a particular dataset.

Another data augmentation method based on data interpolation is *mixup* [20]. In this method, a new sample is generated based on the affine combination of two randomly drawn samples with a random λ drawn from a beta distribution. Computing the loss of a generated sample through this method is a weighted average of the losses for each original label. This relatively new technique has shown to be particularly effective for brain segmentation in MRI exams [9].

6.2.3 Learning-Based Augmentation

A recent development in deep learning technology is the formulation of the generative adversarial neural network (GAN) framework [21–23]. A typical GAN framework consists of a pair of deep convolutional neural networks where one network (generator) generates an image while the other network (discriminator) predicts whether the input is original data in the dataset or synthesized data generated by the generator. In a nutshell, in this framework, two networks are trained to compete with each other and are co-optimized, which results in generating realistic synthesized data. Although this approach requires a large amount of data to train the generator network successfully and a careful training configuration to avoid collapse, using synthesized data from a generator for data augmenting in training deep learning models outperforms training with a combination of other data augmentation techniques [10, 11]. However, some current limitations of using GANs for data augmentation are that synthesized data are restricted to a particular resolution (typically 64 $\times$ 64 or 128 $\times$ 128 pixel) to ensure realistic visual features, and components in a generated image often lose a relative positional consistency [23].

6.2.4 *Practical Considerations*

Of note, there are other data augmentation methods that are not discussed in this chapter, such as random noise addition or lens distortion. Except for applications in skin images, these techniques are empirically less effective for training deep learning models for medical application and are not popular in this domain.

In most major deep learning frameworks, such as PyTorch and TensorFlow, methods listed in this section are either officially supported by the application programming interfaces (APIs) [24, 25] or are implemented and available on public code repositories. Some of the most popular publicly-available libraries for data augmentation are included in the references at end of this chapter [26–28]. Nevertheless, selecting the set of effective augmentation techniques and their optimal parameters for a particular dataset requires a comprehensive understanding of the task, data properties, and the deep learning approach. In general, applying more data augmentation methods potentially makes a model more robust to unseen data samples. In practice, however, only a few of the most effective augmentation techniques are employed in training to reduce computational cost and accelerate convergence during model optimization. As a practical approach to choosing an appropriate data augmentation technique, the effectiveness of a data augmentation technique is assessed based on the similarity of the synthetized data to an unseen validation set. In other words, if an image generated through an augmentation technique resembles the data points with the same class label from a validation set, that augmentation technique would be appropriate for the dataset. Figure 6.1 illustrates the effect of different augmentation techniques on an RGB microscopic image and a gray-scale sagittal CT scan. Some samples exhibit major artifacts (Fig. 6.1d, g, h), which may cause an unfavorable performance degradation depending on the application.

Table 6.1 summarizes the data augmentation techniques based on their methodological category and the dataset modalities that have been utilized. The corresponding references for these techniques are also included in this table. As can be seen in the table, data augmentation techniques are more commonly used for analyzing microscopy images compared to the other modalities.

6.3 Data Augmentation in Medical Applications

In general, deep learning researchers have been successful in developing low bias and low variance models for image analysis. The nonlinear representations of the input data in these models provide robust and accurate models compared to the traditional and shallow machine learning models. The essential requirements of an effective deep learning model include access to a large training dataset, employment of high-performance computing hardware resources such as a graphics processing unit (GPUs), nonlinear representation of the input data, and careful optimization of the parameters and hyperparameters in the model.

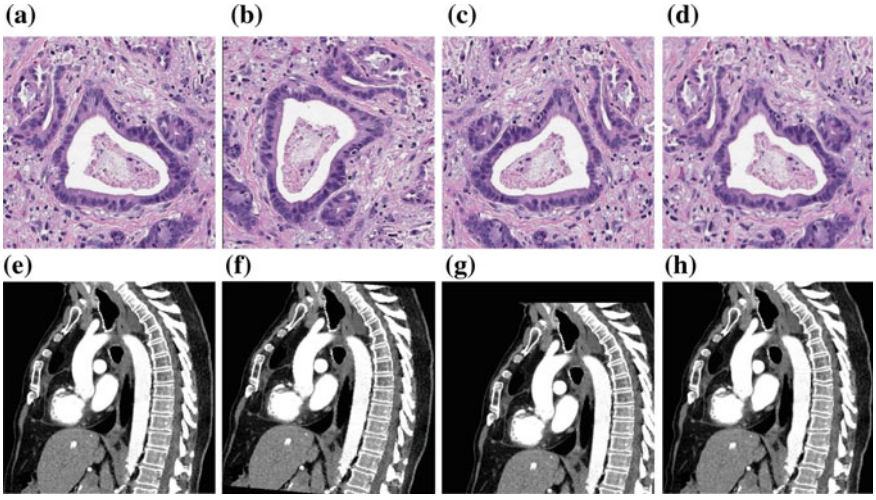


Fig. 6.1 Examples of images generated by different augmentation techniques. Top row: **a** original microscopic image, **b** rotation, **c** horizontal flip, and **d** elastic deformation techniques respectively applied on a microscope image. Bottom row: **e** original, **f** slight rotation, **g** translation, and **h** elastic deformation applied on a slice of CT exam in the sagittal view. The appropriate parameters for these transformations require manual validation to avoid artifacts

In addition to increasing the size of training dataset, there are other approaches that improve the accuracy of deep learning models, such as modifying the objective function or utilizing regularization [39]. Nevertheless, data augmentation is one of the most practical and effective approaches for medical applications to improve the performance of the model. In this section, we focus on the reliance of deep learning models on large training datasets, which can be challenging in medical applications, and the role of data augmentation in tackling this challenge. Furthermore, we review effective augmentation techniques stratified by medical image analysis applications.

6.3.1 Classification

In training a classifier, high variance or overfitting occurs when the error rate on the training set is small, while the error increases on unseen data called test set data. High variance or an overfitted model is defined below, where f is the the classification model:

$$Error_{test}(f) \gg Error_{train}(f)$$

Along with adding a regularization term to the classifier's objective function, data augmentation is another common practical solution to tackle the overfitting problem.

Table 6.1 Summarization of the discussed data augmentation techniques based on their methodology and target modalities. The associated references are included for each case

Augmentation category	Augmentation methods	Microscopy	CT	MRI	Fundus	X-ray	Mammography	Skin images
Basic	Flipping	[29, 30]	–	–	–	[12]	–	[10]
	Scaling	[29, 31]	–	–	[32, 33]	[12, 13]	–	–
	Gaussian NOISE	–	[18]	[18]	–	–	–	–
	Lens correction	–	–	–	–	–	–	[34]
	Rotation	[29–31]	[18, 19, 35]	[18, 36]	[32, 33]	[13]	–	[10]
	Cropping	[29]	[18, 35]	[18]	[33]	[12, 37]	–	–
	Translation	[31]	[18, 35]	[18]	–	[13, 37]	–	–
	Reflection	[31]	–	–	–		–	
	Color	[29–31]	–	–	–	[12]	–	[34]
Interpolation based	Elastic warping	[38]	[18, 19]	[18]	–	–	[17]	[10]
Learning based	GAN	–	[11]	–	–	[37]	–	–

In general, data augmentation adds more samples with different variations in order to avoid overfitting. Therefore, most deep learning classifiers use one or more data augmenting techniques for various medical image modalities.

Generative or GAN-based augmentation methods introduce synthesized samples to improve the classification accuracy. These synthesized samples are optimized for resemblance to the original through a GAN framework. Therefore, GAN-based techniques tend to improve the accuracy of classifiers more than other types of augmentation methods, which transform the images by using random parameters. Developing GAN-based methods for data augmentation is not very common in medical image analysis, mainly due the scarcity of available data. Thus, most published work in this domain is focused on non-medical applications. Few published studies on the use of generative models in medical image analysis are focused on classifying CT scans for liver [11] and chest X-ray images [37]. These experiments evaluated and compared the traditional and generative data augmentation methods for classification and showed that utilizing the generative data augmentation improved the classification accuracy in comparison to basic augmentation.

Spatial transformation of images is a common approach for generating a larger dataset in developing deep classifiers. Choosing the transformation type depends on the dataset and the classifier's objective. For instance, chest X-ray images always have the same orientation, so 90° rotation of a chest X-ray does not generate a valid sample and would even drop the classification accuracy. Furthermore, in some cases, data transformations require a careful parameter selection since they may change the samples' corresponding labels. Applying elastic deformation on bone fracture classification is an example of such transformation [35].

Basic data augmentation methods usually use one or a sequence of valid spatial transformations on medical images. As an example, MRI images have been analyzed for brain tissue classification. Brain tumors appear in different shapes, sizes and locations of the brain in MRI exams. Abnormal tissues such as brain tumors are usually found only on a few slices of an MRI exam. Therefore, spatial data augmentation methods are commonly applied on cropped images of brain tumors to construct a balanced number of samples of tumor and normal classes [36].

Developing CNN-based classifiers on CT, MRI, and X-ray images is based on extracting features from grayscale pixels/voxels, while medical images in other modalities, such as microscopic, fundus, and skin images [34], provide color channel information in addition to the spatial features that are available in the images. Color channel shifting is an additional technique that can be used to augment data in the modalities with colored images.

Microscopic images provide morphological and cell level information of tissues. These high-resolution images are utilized to track diseases that are progressive at the cell level, such as different types of cancers and tumors. Tumor classification is one of the primary goals in analyzing microscopic images. Existing deep learning frameworks require one to upload the images on working memories of GPUs for model training and deployment. Although this requirement significantly enhances the performance of the deep learning models for typical images, it is not feasible to transfer high-resolution microscopic images, which are commonly very large, to

Table 6.2 List of published quantified improvements in deep learning models for medical image classification due to data augmentation

Augmentation methodology	Modality	Accuracy without augmentation (%)	Accuracy with augmentation (%)	Accuracy improvement (%)
Basic augmentation	Colorectal polyps on microscopic images [30]	89.00	91.30	2.30
	Liver CT [11]	57.00	71.60	14.60
	Chest X-ray [37]	81.90	83.12	1.22
GAN based methods	Liver CT [11]	57.00	78.60	21.60
	Chest X-ray [37]	81.90	84.19	2.29

memory. Therefore, the microscopic images are usually cropped for analysis with a feasible memory requirement. Thus, spatial augmentation is applied on cropped images to improve the classification performance [29, 31]. In addition, color data augmentation was recently employed on microscopic images for colorectal polyp classification to improve the accuracy [30].

Fundus images are also colored images, which are used to diagnose eye diseases. Although color augmentation might be useful on retinal images, most of the recently published studies only utilize basic data augmentation methods that do not include color augmentation. For example, a set of basic data augmentation techniques, such as rotation and scaling, are used on fundus images to train a deep learning model to classify diabetic patients, and those techniques outperformed the conventional machine learning methods [33, 40].

For further comparison, in Table 6.2 we listed the published improvements in the classification accuracy of deep learning models as noted in the medical image analysis literature due to data augmentation.

6.3.2 Medical Image Detection

Organ or lesion detection is a key part of medical diagnosis. Typical medical detection tasks consist of localizing small lesions in a full image. The applications of medical image detection range from nodule detection in chest CT images, to lesion detection in mammography, to abnormality detection in retinal images.

Although general object detection frameworks in the field of computer vision align well with medical image detection tasks, there are significant differences in

their training, as medical object detection training commonly involves imbalanced classes in the datasets.

Of note, the number of deep learning models focusing on object detection in medical images is significantly fewer than classification and segmentation models. Therefore, the role of data augmentation for medical image detection has not been fully explored. However, as an example for this task, a deep learning model for breast lesion detection on mammogram images in breast cancer screening has used interpolation and elastic deformation augmentation methods in training to capture different variations of the breast lesions in the images [17].

6.3.3 Medical Image Segmentation

Image segmentation focuses on identifying a similar set of pixels/voxels in a region of interest (ROI). Therefore, medical image segmentation allows quantitative analysis of clinical parameters related to volume and shape. For instance, monitoring brain tumor growth before and after the treatment is one of the vital clinical proceedings that can be performed through tumor segmentation on the corresponding medical images.

Deep learning segmentation techniques on medical images are commonly evaluated by comparing the segmentation results with images manually segmented by clinical professionals. Due to the pixel-level nature of segmentation, preparing labeled data for model training and evaluation is an extremely time consuming and labor-intensive task, and data augmentation methods are critical for developing segmentation methods in this domain.

Of note, organs appear with different shapes, sizes and locations; therefore, the random transformation of each organ can be helpful to add variation to the dataset. Hence, basic augmentation methods for medical image segmentation are usually used to generate more samples for model training.

Spatial and interpolation data augmentation methods are the two augmentation methods that are commonly applied on medical image segmentation. U-net is a popular convolutional neural network which outperforms sliding window CNN segmentation methods [16]. U-net architecture uses elastic deformation, which guarantees better results due to generating more realistic variations for data augmentation. In image segmentation, elastic deformation is also used to avoid overfitting [38].

In practice, basic data augmentation techniques are used to improve segmentation for microscopic images. For example, in developing a fully convolutional neural network for colon gland segmentation, utilizing basic data augmentation methods on microscopic cropped images improved the results [41]. Another study showed that basic data augmentation improved the segmentation of nuclei on histopathology images along with *L2* regularization [31].

Blood vessel segmentation on fundus images has been of great interest to clinicians. Segmenting blood vessels in fundus images is a challenging task because of the small size of the vessels. Blood vessels appear in different angles and widths

across different patients. Therefore, randomly generated images using spatial transformation improve blood vessel segmentation [32]. Of note, the GAN-based models are not employed in analyzing fundus images since providing a ground truth and labeling the vessels are extremely difficult for these images.

6.4 Conclusion

Recent advances in deep learning technology have provided new opportunities for developing and deploying effective medical image analysis models to assist clinicians in medical diagnosis and practice. However, developing such medical image analysis models universally involves one common problem: lack of high-quality labeled training data. This is because acquiring and labeling training samples that have overlapping annotations by multiple domain experts requires enormous resources. Inevitably, the lack of such datasets can slow down the progress and adoption of deep learning for medical image analysis. To address this problem, data augmentation methods can provide a practical solution to expand training datasets and improve the generalizability, stability, and performance of image analysis models, in combination with other model refinements. In this chapter, we have provided an overview of common data augmentation methods and their corresponding characteristics in various medical image analysis applications. We have also provided pieces of practical advice on the appropriate data augmentation techniques to use in different applications and imaging modalities based on the growing body of published literature in the field. This material can be helpful for developing effective deep learning models based on limited amounts of annotated medical images.

References

1. He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
2. Bengio, Y.: Learning deep architectures for AI. Found. Trends® Mach. Learn. **2**(1), 1–127 (2009)
3. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533 (1986)
4. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int J Comput Vision **115**(3), 211–252 (2015)
5. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: Icdar. IEEE (2003)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
8. He, K., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016)
9. Eaton-Rosen, Z., et al.: Improving data augmentation for medical image segmentation (2018)

10. Izadi, S., et al.: Generative adversarial networks to segment skin lesions. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
11. Frid-Adar, M., et al.: Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
12. Larson, D.B., et al.: Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology **287**(1), 313–322 (2017)
13. Lee, H., et al.: Fully automated deep learning system for bone age assessment. J Digit Imaging **30**(4), 427–441 (2017)
14. Fischer, A.H., et al.: Hematoxylin and eosin staining of tissue and cell sections. Cold Spring Harbor Protoc. **2008**(5), pdb. prot4986 (2008)
15. Biberacher, V., et al.: Intra-and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. Neuroimage **142**, 188–197 (2016)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2015)
17. Castro, E., Cardoso, J.S., Pereira, J.C.: Elastic deformations for data augmentation in breast cancer mass detection. In: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE (2018)
18. Christ, P.F., et al.: Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. arXiv:1702.05970 (2017)
19. Sugino, T., et al.: Automatic segmentation of eyeball structures from micro-CT images based on sparse annotation. In: Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging. International Society for Optics and Photonics (2018)
20. Zhang, H., et al.: Mixup: Beyond empirical risk minimization. arXiv:1710.09412 (2017)
21. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)
22. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015)
23. Salimans, T., et al.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (2016)
24. Paszke, A., et al.: PyTorch (2017)
25. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: *OSDI* (2016)
26. https://www.tensorflow.org/api_docs/python/tf/contrib/image.
27. https://opencv.org.
28. https://github.com/aleju/imgaug.
29. arXiv:1606.00897Bauer, S., et al.: Multi-organ cancer classification and survival analysis. (2016)
30. Korbar, B., et al.: Deep learning for classification of colorectal polyps on whole-slide images. J. Pathol. Inf. **8** (2017)
31. Veta, M., Van Diest, P.J., Pluim, J.P.: Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2016)
32. Maninis, K.-K., et al.: Deep retinal image understanding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2016)
33. Yang, Y., et al.: Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2017)
34. Galdran, A., et al.: Data-driven color augmentation techniques for deep skin image analysis. arXiv:1703.03702 (2017)
35. Tomita, N., et al.: Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans.**98**, 8–15 (2018)
36. Pereira, S., et al.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging **35**(5), 1240–1251 (2016)

37. Madani, A., et al.: Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: Medical Imaging 2018: Image Processing. International Society for Optics and Photonics (2018)
38. Quan, T.M., Hildebrand, D.G., Jeong, W.-K.: Fusionnet: a deep fully residual convolutional neural network for image segmentation in connectomics (2016)
39. Goodfellow, I., et al.: Deep learning, vol. 1. MIT press, Cambridge (2016).
40. Abràmoff, M.D., et al.: Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci **57**(13), 5200–5206 (2016)
41. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2016)

Part III
Medical Applications and Reviews

Chapter 7
Application of Convolutional Neural Networks in Gastrointestinal and Liver Cancer Images: A Systematic Review

Samy A. Azer

Abstract *Background*: The use of artificial intelligence in the interpretation of images and the diagnosis of gastrointestinal and liver cancers has been evaluated. A convolutional neural network (CNN), a machine-learning algorithm similar to deep learning, has shown its capability to recognize specific features that can detect cancer. *Aims*: This review aimed at assessing the application of CNN in examining gastrointestinal and liver images in the diagnosis of cancer and explore the accuracy level of CNNs used. *Methods*: PubMed, EMBASE, and the Web of Science were systematically searched. Studies using CNNs to analyze endoscopic, pathological, or radiological images of gastroenterological or liver cancers were identified according to the international consensus standards with the aim of detecting or staging cancer. Two independent reviewers extracted the data for the study reports. The accuracy of CNNs in detecting cancer or early stages of cancer was analyzed. The primary outcomes of the review were analyzing the type of cancer, and identifying the type of images that showed optimum accuracy in cancer detection. *Results*: A total of 22 articles that met the selection criteria and were consistent with the aims of the study were identified. The studies covered cancers of the esophagus, stomach, pancreas, liver and biliary system and colon. It also covered risk factors and pre-cancerous conditions such as *Helicobacter pylori* infection, liver cirrhosis and colonic polyps. The studies were performed in Japan (n = 6), China (n = 6), the United States (n = 5), and Hong Kong, France, Switzerland, Germany, the United Kingdom, and Bangladesh (n = 1 each). The studies aimed at identifying lesions (n = 5), classification (n = 9), and segmentation (n = 8). Several methods were used to assess accuracy of the CNN and the overall level was satisfactory. *Conclusions*: The role of CNNs in analyzing images and as tools in early detection of gastrointestinal or liver cancers and classifying cancers has been demonstrated in these studies. Although a few limitations have been identified in these studies, overall there was an optimal

S. A. Azer (✉)
Gastroenterologist and Professor of Medical Education, Chair of the Curriculum Development and Research Unit, Department of Medical Education, College of Medicine, King Saud University, P O Box 2925, Riyadh 11461, Saudi Arabia
e-mail: azer2000@optusnet.com.au

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_7

level of accuracy of the CNNs used in segmentation and classification of images of gastrointestinal and liver cancers.

Keywords Deep learning · Convolutional neural network (CNN) · Gastrointestinal cancers · Liver cancer · Medical imaging · Classification · Segmentation · Machine learning · Artificial intelligence · Computer-aided diagnosis

7.1 Introduction

A progressive interest in the use of deep learning has recently emerged, particularly the use of convolutional neural networks (CNNs), a class of artificial neural networks that have been widely used in biomedical and clinical research [1]. For example, the potential use of CNNs has been shown in diabetic retinopathy screening [2], skin lesion classification [3], and interpretation of blood culture [4]. There is also a surge of interest in the potential of CNNs in radiology research [1, 5], screening for cancer via endoscopic examinations [6], and pathology research [7]. Moreover, several studies have demonstrated the usefulness of CNNs in oncology: *Lesion detection*, in which endoscopists, radiologists, and pathologists detect abnormalities using medical images has been aided by CNNs, including the detection of colonic polyps, liver cancer on radiological images, and histopathological malignant changes in biopsy specimens [8]. CNNs can also aid in tumour *Classification,* wherein deep learning utilizes target lesions depicted in medical images to classify lesions into two or more classes, e.g., lesions/non-lesions or malignant/benign. Other examples include classification of histopathological images or the classification of lung nodules on computed tomography (CT) [9]. Therefore, the task is to determine "optimal" boundaries for separating classes in the multi-dimensional feature space that is formed by input features. CNNs also have potential use in *Segmentation* of organs or anatomical structures and *image reconstruction*, which may include obtaining a noiseless CT image reconstructed from a subsampled sonogram [10]. Segmentation is a functional image processing technique for the analysis of medical images, such as quantitative evaluation of clinical parameters and computer-aided diagnosis system [11]. With this information in mind, this chapter aims at reviewing and identifying the applications and uses of CNN in the interpretation of gastrointestinal and liver cancer images.

Gastrointestinal cancers comprise colorectal cancer, the third most prevalent cancer in the world in both men and women; pancreatic cancer, the fourth most prevalent cancer in both men and women; and liver and biliary tract cancers, the fifth most prevalent cancer in men and the eighth in women [12]. Moreover, cancers of the pancreas and liver have the lowest survival rates of all cancer types (8% and 18%, respectively) [12]. These cancers with high prevalence and poor prognosis require more research attention, and early detection and clinical intervention may improve survival rates. Interestingly, colorectal cancer incidence patterns were generally similar in men and women from 2005 to 2014, and the incidence rates declined annually

by about 2–3% because colonoscopy among US adults aged 50 years and older tripled from 21% in 2000 to 60% in 2015 [13], indicating the significance of early detection of colonic polyps and premalignant lesions in reducing the rates and prognosis of colorectal cancer [14].

We would like to indicate here that the prevalence of precancerous polyps in the population older than 50 years is approximately 50% [15]. Adenomas are the most prevalent precancerous polyps, and use the adenoma detection rate, or the percentage of screening colonoscopies with more than one adenoma detected [16, 17], to assess the ability of colonoscopists to find adenomas during examination. However, the adenoma detection rate varies widely, between 7 to 57%, among colonoscopists performing these procedure [18], indicating variability in adenoma detection. In addition to the training skills of the colonoscopist, several other factors may affect the adenoma detection rate, including the quality of patient's preparation for colonoscopy, the duration of the examination, and the technique used [19, 20]. A larger study recently found that, for each 1% increase in the detection of adenoma, the colorectal cancer rate was decreased by 3% [18]. This indicates that there is a need to optimize the adenoma detection rate and minimize any failure to diagnose polyps and adenomas during screening.

The use of deep learning by using CNNs may offer a solution to this challenge.

Other gastrointestinal cancers that can be detected early include gastric cancer, through the early detection and treatment of *Helicobacter pylori* infection of gastric mucosa [21], and the detection of Barrett's oesophagus (esophageal adenocarcinoma) [22].

Hepatocellular carcinoma (HCC), also known as hepatoma, accounts for more than 90% of all cases of primary liver cancer. It is the six most common type of cancer worldwide and its incidence has increased significantly over the last two decades, leading it to become the third leading cause of cancer-related mortality [23]. The burden of HCC varies depending on geographical location and it is a major public health problem in the Asia-Pacific region [24].

The aim of this review was to evaluate the use of CNNs in images of gastrointestinal and liver cancers. We did this by assessing the current status of CNNs and their applications in gastrointestinal and liver cancers, identifying gaps and deficiencies in research in this area, particularly in relation to gastrointestinal and liver cancers, and determining future directions in this area and research priorities to maximise the applications of research in gastroenterology cancers. Therefore, our research questions are: (i) what is the current status of research output in the use of CNNs in assessing gastrointestinal and liver cancer images? and, (ii) what is the accuracy of CNNs/deep learning systems, in lesion detection, classification, or segmentation of these images?

7.2 Methods

This systematic review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [25].

7.2.1 Study Selection

The PubMed, EMBASE, and the Web of Science databases were searched for studies on CNNs in gastrointestinal and liver cancer. The reasons for selecting these databases were to avoid publication bias (for example, geographical bias or bias against publications of negative results), easy user interface, and maximization of yield. The search covered studies published up to May 2018. The search was limited to studies in English and conducted in human patients. Studies on animals or animal models were not included. We searched the databases for articles with contributions of the subject headings and the following key words: "Cancer", "Esophagus", "Stomach", "Gastric", "Pancreas", "Liver", "Hepatic", "Biliary system", "Gallbladder", "Colon", "Polyp", "Adenoma", "Gastrointestinal", "Pathology", "Histology", "Histopathology", "Malignancy", "Dysplasia", "Metaplasia". All search was performed independently by two researchers (SAA, and SA). To maximize the search yield, another search was performed manually by searching the reference lists of the primary articles and reviews to identify studies not found by the database search [26].

We also searched the journals listed by the Journal Citation Reports-2016 of the Web of Science under the category gastroenterology and hepatology, oncology, computer sciences and engineering, and medical informatics.

7.2.2 Criteria for Consideration of Studies

To identify targeted studies, we created a PICOS framework (Population, Intervention, Comparison, Outcome, Studies) for inclusion and exclusion. Table 7.1 sum-

Table 7.1 PICOS framework (population, intervention, comparison, outcome, studies) to identify studies for inclusion

Population	Worldwide, datasets, men and women, humans
Intervention	Use of CNN-diagnostic programs
Comparison	Normal population, use of manual detection by colonoscopists, or pathologists examining images/slides
Outcome	Lesion detection, classification, segmentation, or image reconstruction
Studies	Controlled, or comparison to manual (routine) assessment

marizes the PICOS framework used. Studies that reported data on the use of CNNs in gastrointestinal and liver cancer images were included. The search was limited to studies in the English language and conducted in humans. Studies on animals or animal models were not included. Reviews, editorials, commentaries, letters to the Editors, conference proceedings, abstracts, or monographs were not included.

7.2.3 Study Selection

The two researchers (SAA and SA) independently reviewed the titles and abstracts of all citations identified by the literature search. Relevant studies were retrieved and reviewed in detail. Any disagreement was discussed by the two evaluators until they came to a consensus. The full texts of potentially relevant articles were sought and the selection criteria were applied. Reviewers were not blinded to authors' names or institutions. Studies were selected if they match the selection criteria.

7.2.4 Data Extraction

Data were extracted independently by two researchers using a predefined extraction form. The following data were abstracted in the form: (1) first author's name, (2) year of publication, (3) objectives/research question, (4) method used, (5) gastrointestinal cancer investigated, (6) main results, (7) accuracy, sensitivity, specificity of method used, and (8) institute, university, city and country where the study was conducted. Details on reported statistical associations and comparison of the results obtained with those obtained by using other methods were also evaluated.

Agreement between evaluators measured by the degree of inter-rater agreement using the Cohen kappa coefficient was also carried out using SPSS (version 24, SPSS statistics/IBM Corp. Chicago IL, USA) software [27].

7.3 Results

7.3.1 Literature Search and Selection Process

Figure 7.1 is a flow diagram summarizing the search and selection process of articles in the literature. One hundred ninety-two potentially relevant publications were identified through the search of the databases. After removal of duplicates, 120 articles remained. Of these, 79 did not fit the inclusion criteria. Forty-one full-text articles were assessed for eligibility. Finally, we identified 22 articles that met our selection criteria and were consistent with the aims of the systematic review [28–49].

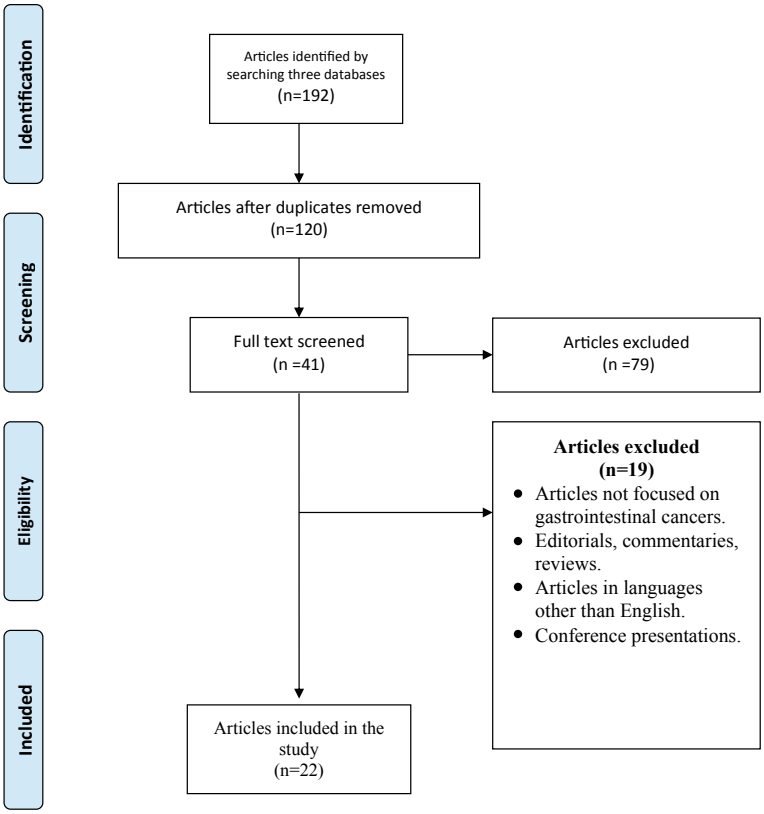


Fig. 7.1 A PRISMA flowchart showing articles searched on use of convolutional neural networks in gastrointestinal and liver cancers images

7.3.2 Characteristics of Included Studies

Table 7.2 summarizes details of the 22 studies included [28–49]. The studies analyzed images of cancers of the following anatomical organs: Esophagus [28–30], stomach [31–34], pancreas [35, 36], liver and biliary system [37–41], and colon [42–49]. The studies also covered risk factors and possible predisposing causes/conditions, such as *H. pylori* infection, which could predispose to gastric cancer [31, 32], liver cirrhosis, which can predispose to HCC [38], and colonic polyps that could predispose to colon cancer [42–44]. Although these studies examined gastroscopy images [28, 32–34], colposcopy images [42–45], ultrasound images [38], CT images [37, 40, 49], and magnetic resonance imaging (MRI) images [35, 39], other images, such as cellular and histopathological images, were also included [36, 41, 46–48].

Table 7.2 Summarizes the role of Convolutional Neural Networks (CNNs) in the diagnosis and management of gastrointestinal and liver disorders

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Takiyama et al. (2018) [28]	Can CNN recognize the anatomical location of esophagogastroduodenoscopy (EGD) images in an appropriate manner?	A CNN-based diagnostic program was constructed based on GoogLeNet architecture and was trained with 27,335 EGD images categorized into four anatomical locations	The trained CNN showed robust performance in its ability to recognize the anatomical location of EGD images	Receiver operating characteristics (ROC) curves showed high performance of the trained CNN to classify the anatomical location of images	Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama (Japan)
Fechter et al. (2017) [29]	The authors proposed random walker approach driven by a 3D fully convolutional neural network (CNN) to automatically segment the esophagus from CT images	An active contour model (ACM) is filtered to a CNN soft probability map to get first estimate of the esophageal location The output of the CNN and ACM are then used to drive random walker	The developed CNN model yielded accurate estimates of esophageal location	The method outperformed all existing approaches	University of Freiburg German Cancer Consortium (DKTK), Heidelberg (Germany)
Xue et al. (2016) [30]	Extraction and the classification of microvascular morphological type to aid esophageal cancer detection	A specialized CNN is designed to extract hierarchical features and Support Vector Machines (SVM) are used to enhance the generalization ability of classifiers	The system was able to assist clinical diagnosis to a certain extent	The recognition rate was 88.2% on patch level	University of Science and Technology of China, Hefei (China)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Shichijo et al. (2017) [31]	Evaluate the role of CNN in the diagnosis of *H. pylori* gastritis based on endoscopic images	Deep CNN, pre-trained and fine tuned on database of 32,208 images either positive or negative for *H. pylori* were used Another CNN was trained using images classified according to 8 anatomical locations (Secondary CNN) GoogleNet architecture at https://arxiv.org/abs/1409.4842 Fine-tuned by using Adam at https://arxiv.org/abs/1412.6980	H pylori gastritis could be diagnosed on endoscopic images using CNN with high accuracy and in considerably shorter time compared to manual diagnosis	The sensitivity, specificity, accuracy and diagnostic time were 81.9%, 83.4%, 83.1% and 198 s, respectively For secondary CNN, the results were 88.9%, 87.4%, 87.7%, and 194 s, respectively	Osaka International Cancer Institute, Osaka (Japan)
Itoh et al. (2017) [32]	Develop CNN, a machine learning algorithm, similar to deep learning and capable of recognizing specific features of gastric endoscopy images to detect *H pylori* infection at an earlier stage	CNN was developed using 179 upper gastrointestinal endoscopy images obtained from 139 patients	CNN-aided in the diagnosis of *H pylori* is feasible and is expected to facilitate improvement in early detection of *H pylori* infection	The sensitivity and specificity for the detection of H pylori were 86.7% and 86.7% respectively and the AUC was 0.956	Graduate School of Engineering, Chiba University, (Japan)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Hirasawa et al. (2018) [33]	Assess the ability of a developed CNN to automatically detect gastric cancer in endoscopic images	The CNN-based diagnostic system was constructed based on Single Shot MultiBox Detector architecture and trained using 13,584 endoscopic images of gastric cancer	The CNN required 47 s to analyze 2296 test images. The CNN correctly diagnosed 71 of 77 gastric cancer lesions All missed lesions were difficult to distinguish from gastritis even for experienced endoscopists	Overall sensitivity of diagnosing gastric cancer was 92.2% and the positive predictive value was 30.6%	Japanese Foundation for Cancer Research, Tokyo (Japan)
Zhang et al. (2017) [34]	Assess the ability of CNN in detecting gastric precancerous disease (GPD)	CNN with a concise model called the Gastric Precancerous Disease Network (GPDNet) was used Dataset repository at https://github.com/jiquan/Data-Open-Acess4Plos.Onegit	The system is able to accurately detect gastric precancerous disease	The final accuracy of GPD Net was 88.9%, which is promising for clinical diagnosis for gastric precancerous disease recognition	Zhejiang University, Hangzhou (China)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Cai et al. (2016) [35]	Can we conduct pancreatic detection and boundary segmentation using CNN	Two types of CNN models were used: (1) tissue detection step, and (2) boundary detection step to locate the boundaries of pancreas	The proposed algorithm achieved the best results in detecting and identifying the boundaries of the pancreas	The mean dice similarity coefficient (DSC) was 76.1% with a standard deviation of 8.7% in a dataset containing 78 abdominal MRI scans	University of Florida, Gainesville, Florida (The United States)
Xing et al. (2016) [36]	Determine the use of CNN and computer-aided images analysis in the early detection of pancreatic neuroendocrine tumour (NET)	A deep CNN model to generate a probability map was used	Using a novel segmentation algorithm to separate individual nuclei combining a robust selection-based sparse shaped model and a local repulsive deformable model in the detection of pancreatic neuroendocrine tumour (NET)	The proposed algorithm has been tested on three large-scale pathology image databases using a range of different tissues and stain preparations, and compared to other state of the arts approaches, this algorithm demonstrates superior outcomes	University of Florida, Grainesville, Florida (The United States)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Liu et al. (2018) [37]	Determine the accuracy of segmentation of specific organs such as the liver on computed tomography (CT) scans by using convolutional neural networks (CNNs)	Automatic organ segmentation for CT scans by using convolutional neural network (CNNs) and Support Vector Machine (SVM) Open Chinese dataset TIANCHI Medical AI challenge at https://tianchi.aliyun.com/competition/	The proposed method can precisely and efficiently detect liver segmentation	The Dice Coefficient of the liver segmentation results reached to 97.4%	Jilin University, Jilin (China)
Liu, et al. (2017) [38]	The authors proposed computer-aided cirrhosis diagnosis system to diagnose cirrhosis based on ultrasound images and a deep convolutional neural network (CNN) model	Using a CNN mode to extract features from ultrasound images and finally a trained Support Vector Machine (SVM) classifier is applied	The CNN model can identify cirrhosis and classify ultrasound images with high accuracy	The proposed model can effectively extract the liver capsule and accurately classify the ultrasound images	Fudan University, Shanghai (China)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Ibragimov et al. (2017) [39]	The authors proposed a novel framework for automated segmentation of the portal vein (PV) from computed tomography (CT) images using CNN	Convolutional neural networks (CNNs) have been used to learn the consistent appearance pattern of the PV using a training set of CT images. Markov random fields (MRFs) were further used to enhance the CNN work and remove mis-segmental regions	The CNNs and anatomical analysis have shown the ability to accurately segment the PV and can be potentially integrated into liver radiation therapy planning	The obtained accuracy of the segmentation was 0.83 and 1.08 mm in term of the median Disc Coefficient and mean Symmetric Surface Distance, respectively	Stanford University School of Medicine, California (The United States)
Yasaka et al. (2018) [40]	Investigate the diagnostic performance by using deep learning method with a convolutional neural network (CNN) for differentiating liver mass at dynamic contrast agent –enhanced computed tomography	CT images of liver masses over three phases were used and the CNN composed of six convolutional were used	Deep learning with CNN showed high diagnosis performance in differentiation of liver masses at dynamic CT	The median accuracy of differential diagnosis liver masses for test data was 0.84 Median area under the receiver operating characteristic curve was 0.92	University of Tokyo Hospital, Tokyo (Japan)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Qin et al. (2018) [41]	Develop a novel system for automated liver segmentation	A novel superpixel-based and boundary sensitive convolutional neural network (SBBS-CNN) pipeline was used for liver segmentation Dataset at https://competitions.codalab.org.competitions/17094	The SBBS-CNN model provided an accurate and effective tool for automated liver segmentation	The model showed superior performance in comparison with state-of-art methods, including U-Net, pixel-based CNN, active contour, level-sets and graph-cut algorithms The SBBS-CNN achieved mean Dice Similarity Coefficients of 97.3% and the average Symmetric Surface Distance was 1.77 mm	Clinical Academy of Sciences, Shenzhen (China) and Stanford University, Palo Alto, California (The United States)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Urban et al. (2018) [42]	Test the ability of computer-assisted image analysis with CNN to improve polyp detection	CNN tested 20 colonoscopy videos, a total of 5 h Video available at: https://www.igb.uci.edu/colonoscopy/AI_for_GI.html Supplementary material at https://doi.org/10.1053/j.gastro.2018.06.037	Four expert reviewers of the videos identified 8 additional polyps without CNN assistance that had not been removed and identified additional 17 polyps with CNN assistance	The CNN identified polyps with an area under the receiver operating characteristic curve of 0.991 and an accuracy of 96.4% The CNN had a false positive rate of 7%	University of California, California (The United States)
Billah et al. (2017) [43]	Can an automated system support in gastrointestinal polyp detection?	CNN combined with a linear Support Vector Machine (SVM) Most of the data were collected from Department of Electronics, University of Alcala at https://www.depeca.uah.es/colonoscopy-dataset/ Also, Endoscopic Vision Challenge at https://polyp.grand-challenge.org.databases	The computer-aided polyp detection had reduced the rate of missing polyps and assisted in finding colonic regions to pay attention to	The proposed system outperforms the state-of-the-art methods, gaining accuracy of 98.6%, sensitivity of 98.8%, and specificity of 98.5%	Mawlana Bhashani Science & Technology University, Tangall (Bangladesh)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Zhang et al. (2016) [44]	Developing a fully automatic algorithm to detect and classify hyperplastic and adenomatous colorectal polyps	A novel transfer learning application is proposed utilizing features learned from big non-medical data sets with 1.4–2.5 million images using deep Convolutional Neural Network	The method identified polyp images from non-polyp images in the beginning followed by predicting the polyp histology Automated algorithms can assist endoscopists in identifying polyps that are adenomatous but have been incorrectly judged as hyperplasia	Compared with visual inspection by endoscopists, the results of the study shows that the method has similar precision (87.3% versus 86.4%), but a higher recall rate (87.6% versus 77.0%) and a higher accuracy (85.9% versus 74.3%)	The Chinese University of Hong Kong, Shatin (Hong Kong)
Komeda et al. (2017) [45]	Can a computer-aided diagnosis (CAD) based on CNN enable the diagnosis of colonic polyps?	Convolutional Neural Network (CNN) with a computer-aided diagnosis (CAD) system and artificial intelligence used to study endoscopic images	The decision by the CNN was correct in 7 of 10 cases	The system may be useful for rapid diagnosis of colorectal polyp classification Further studies are needed to confirm the effectiveness of a CNN-CAD system in routine colonoscopy	Kindai University Faculty of Medicine Osaka-Sayama (Japan)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Haj-Hassan et al. (2017) [46]	Can CNN predict tissue types related to colorectal cancer progression?	CNN and multispectral biopsy images of 30 patients with colorectal cancer images at three different histopathological stages	CNN has demonstrated the ability to detect colorectal cancer types of tissues with accuracy	The accuracy was 99.2%; outperforming existing approaches based on traditional features extraction and classification techniques	University of Lorraine, Metz, Lorraine (France)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Kainz et al. (2016) [47]	Assess the ability of deep learning for segmentation of glands and classification to differentiate between benign and malignant tissues of the colon	Deep neural-based approach designed for segmentation and classification of glands in colonic tissues into benign or malignant Dataset publicly available at https://www.warwick.ac.uk/bialab/GlasContest Trained models of the deep convolution net available at https://github.com/pkainz/glandsegmentation-models Input data and evaluation scripts are available at https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/glascontest	The model demonstrated the ability to differentiate between benign and malignant colonic tissues with high accuracy	The segmentation performance and tissue classification accuracy has been 98%, and 95%, respectively	University of Zurich, ETH Zurich, Zurich (Switzerland)

(continued)

Table 7.2 (continued)

Author (year) [reference]	Study goal/research question	Method used/dataset links/other links	Main findings	Accuracy method used	University. Institute, City (Country)
Sirinlukunwattana et al. (2016) [48]	Detection and classification of histopathology images of colorectal cancerous tissues among locality sensitive deep learning	A Spatiality Constrained Convolutional Neural Network (SC-CNN) was used to perform nucleus detection and for classification proposed a Novel Neighbouring Ensemble Predictor (NEP) coupled with CNN	On evaluation on a large dataset of colorectal adenocarcinoma images, consisting of more than 20,000 annotated-nuclei belonging to four different classes	The method produced the highest average F1 score as compared to other recently published approaches	University of Warwick, (The United Kingdom)
Men et al. (2017) [49]	Propose a novel Deep Dilated Convolutional Neural Networks (DD CNN)-based method for fast and consistent auto-segmentation in colorectal cancer	Deep Dilated Convolutional Neural Network (DD CNN)	Deep Dilated Convolutional Neural Networks (DD CNN) can be used with accuracy and efficiency to contouring and streamline radiotherapy	The proposal outperformed U-Net for all segmentations The average Dice similarity coefficient (DSC) was 3.8% higher than that of U-Net	National Cancer Center Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing (China)

CAD = Computer-Aided Diagnosis; CNN = Convolutional Neural Network; DD CNN = Deep Dilated Convolutional Neural Networks; DSC = Dice Similarity Coefficient; EGD = Esophagogastroduodenoscopy; SBBS-CNN = Superpixel-Based and Boundary Sensitive Convolutional Neural Network; ROC = Receiver operating characteristics; SVM = Support Vector Machine

7.3.3 Countries and Institutes/Universities Involved

Geographically, these studies were performed in Japan [28, 31–33, 40, 45], China [30, 34, 37, 38, 41, 49], the United States [35, 36, 39, 41, 42], Hong Kong [44], France [46], Switzerland [47], Germany [29], the United Kingdom [48], and Bangladesh [43].

Top universities and research institutes that lead such research included: Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama; Osaka International Cancer Institute, Osaka; Graduate School of Engineering, Chiba University; Japanese Foundation for Cancer Research, Tokyo; University of Tokyo Hospital, Tokyo; Kindai University Faculty of Medicine, Osaka-Sayama; and University of Science and Technology of China, Hefei; Zhejiang University, Hangzhou; Jilin University, Jilin; Fudan University, Shanghai; Clinical Academy of Sciences, Shenzhen; National Cancer Center Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing; University of Florida, Gainesville, Florida; Stanford University School of Medicine, California; Stanford University, Palo Alta, California; The Chinese University of Hong Kong, Shatin; University of Lorraine, Metz, Lorraine; University of Zurich, ETH Zurich; University of Freiburg German Cancer Consortium (DKTK), Heidelberg; University of Warwick; Warwick; and Mawlana Bhashani Science & Technology University, Tangall.

7.3.4 Methods Used

The descriptions of the methods used in these studies varied significantly, and the focus of the content was affected by the authors' backgrounds. Studies that were primarily performed by clinicians focused on the number of patients, number of images and image selection, the endoscopy/radiology procedures used, image preparation for training set, and image preparation for the validation set, and only a brief description of the CNN algorithm used was included [for example, 28, 31, 40, 45]. In contrast, studies that were performed by engineering, computer sciences, or medical informatics departments, with no medical background, provided more detail about the dataset, the generation of the CNN algorithm, the architecture, the Support Vector Machine, and the experiment, and they addressed the technical component in more detail [for example, 30, 36]. Usually, in the first group, the number of patients included and the images used were in the thousands. For example, in one study, the number of patients was 1750, and the authors included 70,000 esophagogastroduodenoscopy images and selected 27,335 images [28]. In the study by Shichijo et al. [31], a total of 32,208 images from patients who were classified as *H pylori*-positive (735 patients) or negative (1015 patients) were prepared for the development of the dataset. In the study by Hirasawa et al. [33], a total of 13,584 endoscopic images of gastric cancer were used and, to evaluate the diagnostic accuracy, an independent test set of 2296 stomach images collected from 69 consecutive patients with 77 gastric cancer lesions was applied to the constructed CNN. In contrast, in a study not authored by clinicians,

a local dataset containing 261 full-size images from 67 patients was used [30]. However, not all studies that were performed by a team representing both groups (medical and computer sciences) had a balanced methodology or a larger number of patients and images. While this may be related to the distribution of tasks and responsibilities among the authors, it is also possibly related to the discipline/specialization of the journal that published the work [for example, 29, 32, 48].

Table 7.2 summarizes the CNN deep learning methods used in these studies. The methods aimed at lesion detection/diagnosis [31–33], localization and identification of polyps [42, 43], classification [28, 30, 34, 38, 44–48], differentiation of liver masses [40], segmentation of the pancreas in MRI [35], segmentation of nucleus [36], organ segmentation [37], segmentation of the portal vein [39], segmentation of colonic glands in stained histopathological sections [47], segmentation of the esophagus in CT [29], segmentation of liver in abdominal CT for radiation therapy planning [41], and segmentation in colorectal cancer [49].

In these studies, the lesions in an image were segmented by use of segmentation technique such as weighted total variation [47], an edge-based segmentation, graph-based data fusion model [35] selection-based sparse shape model and a local repulsive deformable model [36], or simple linear iterative clustering [37]. Features such as contrast, circularity, and size were extracted from the segmentation lesions by use of a feature extractor. In some studies, the extracted features were entered as an input using a linear or quadratic discriminant analysis and a supported vector machine [30, 37, 38, 43].

7.3.5 Accuracy Measures Used

Table 7.2 summarizes accuracy measures used. In these studies accuracy was measured using several approaches, which can be summarised as follows: (i) generating receiver operating characteristic (ROC) curves [28, 40, 42], (ii) comparing the predicted outcomes with the results reported in the literature [29, 41, 48], (iii) measuring sensitivity, specificity, accuracy and diagnostic time [31–33, 43, 46], (iv) measuring the mean dice similarity coefficient [35], (v) measuring accuracy of segmentation, disc coefficient, and symmetric surface distance [37, 39, 47, 49], and (vi) comparing the outcomes with visual inspection by endoscopists (precision and recall rates) [44]. Some studies did not measure accuracy [for example, 45].

7.3.6 The Agreement Between the Evaluators

The inter-rater agreement between evaluators had an overall κ scores in the range of 0.789–0.956.

7.4 Discussion

The aim of this study was to assess the use of CNNs in deep learning of gastrointestinal and liver cancer images. In this review, 22 studies were identified, and they covered common cancers affecting the gastrointestinal system and the liver. Because of the heterogeneous data and the variability in the design of methods and reported results, it was not possible to conduct a meta-analysis. Nevertheless, certain parameters were evident from these studies.

First, the studies have shown the ability and accuracy of CNNs in analysing and offering a diagnosis (segmentation, and classification) from endoscopy images, and radiology images, as well as histopathology and cellular images of gastrointestinal and liver cancers. Although the number of full original articles on each cancer was small, all 22 studies were published between 2016 to 2018, reflecting the fact that deep learning and machine learning as outlined in CNNs and their applications in medical sciences is a recently developed discipline [50].

Second, while the technical information and a description of the test data preparation, the training and evaluation algorithm development, and the methods used in assessing accuracy are vital [51], information about the sources of images, the patients involved in the study, and the clinical information collected is equally important. In this review we noted variability in the methods section in addressing these two components, with no standardization. While these differences might be related to differences in the focus of journals that published these studies and the guidelines provided to authors [52], there is a need for more articles that equally address both aspects within the methods section.

Third, using CNN deep learning in segmentation, classification, and lesion detection in medical images requires such a large number of parameters that they cannot be determined manually but must be retrieved from the data [50, 53]. Therefore, detection and diagnosis tasks in medical imaging require learning from examples or data for determination of this large number of parameters in a complex model. Therefore, CNNs require many training images (e.g., 1,000,000 images) because of the large number of parameters in the model, whereas in other models such as a massive-training artificial neural network (MTANN), a small number of training images may be needed [50, 53]. It is also important as we assess accuracy and performance to compare the performance of models, such as MTANN, with the performance of CNNs under the same conditions. The comparison should include the level of sensitivity and the yielded area under the ROC curve produced on using both models [54].

It would be of interest to assess the performance of different articles that were validated on the same datasets; however, most studies did not provide enough information about the exact source of their datasets and it was therefore not possible to trace which studies used the same testing protocol. Such information is vital for comparison and assessment. We hope that journals interested in publishing such studies on deep learning, CNNs and artificial intelligence develop standardized guidelines

that require authors to state such information, including details about the sources of datasets and protocols used in testing.

Fourth, while writing the discussion of this manuscript we came across a recently accepted manuscript [55]. The study does not exactly fit with our inclusion criteria as it does not solely focus on gastrointestinal cancers, but uses publicly available datasets, including the lymphoma dataset, the breast grading carcinoma dataset, the laryngeal dataset, and the colorectal dataset. However, we thought that it might be useful to compare the bioimage classification used with those included in this study. Table 7.3 summarizes key findings in the paper by Nanni et al. [55] and those identified from the studies included here in relation to colorectal cancer. While we found some useful information to researchers interested in this area, it also highlights gaps in the reported studies and, as stated earlier, there is a need to include more information about the dataset used.

This study, however, is not free from limitations. Considering the diversity of gastrointestinal cancers included, we must interpret the findings with caution, particularly with the lack of specific information needed to conduct meta-analysis [56]. Moreover, there could be publication bias that precluded the publication of negative studies [57]. We only included studies published in English, and there could be high-quality articles related to this topic published in other languages. Additionally, the included studies varied in terms of patient type, study design, disease pathology, disease severity, and details of methods used.

7.4.1 Future Research Directions

This study highlights several future directions for using CNNs to interpret gastrointestinal and liver cancer images. These can be summarized as follows:

First, there is a need for multi-institute, multicenter studies with a large number of patients in each of the gastrointestinal cancers. Such collaborations could resolve the concern about the insufficient amount of training data in the medical image domain. Several researchers reported a smaller size because of training data or difficulties in obtaining images, which make the direct application of machine learning algorithms inappropriate for medical datasets and hence affect the capacity to conduct image classification or image segmentation with high accuracy. This is particularly important in diseases with low incidence or diseases not frequently studied or limited to particular geographic regions, which could make the collection process harder and costlier, since the number of images obtained depends on disease incidence. Again, global collaboration could resolve such challenges [58].

Second, we need case control studies where the use of CNNs can be compared with manual assessment of images by endoscopists, experts and pathologists. One of the major challenges we face with the use of CNNs is the difficulty in choosing discriminant features to represent clinical characteristics and using them as the key features in the CNN algorithm in segmentation and classification functions.

Table 7.3 Comparing the studies on colorectal cancer included in this systematic review with the paper published recently by Nanni et al. [55]

Authors, (year) [reference]	Class	Samples/images	Image size	URL for downloads/links	PI architecture
Nanni et al. (2018) [55]	8	5000	150 × 150	Zenodo.org/record/53,169#.WaXjW8hJaUm	–
URBAN et al. (2018) [42]	–	8,641 images from 2,000 patients (4,088 images of unique polyps and 4553 images without polyps) 44,947 image frames from 9 videos Colonoscopy images resolution 1280 × 1024 pixels	224 × 224	https://www.igb.uci.edu/colonoscopy/AI_for_GI.html	VGG16, VGG19, and ResNet50
Billah et al. (2017) [43]	–	100 videos	227 × 227	–	Three layers of architectures are described
Zhang et al. (2017) [44]	–	1.4–2.5 million images 1104 endoscopic nonpolyp images 826 NBI endoscopic polyp images	227 × 227	–	–
Haj-Hassan et al. (2017) [46]	–	Input image 60 × 60 × 16	512 × 512	–	Three layers of architectures are described

(continued)

Table 7.3 (continued)

Authors, (year) [reference]	Class	Samples/images	Image size	URL for downloads/links	PI architecture
Kainz et al. (2017) [47]	–	165 images of benign and malignant colorectal cancer	Pixel resolution of the images was isotropic at 0.62 μm	https://www.warwick.ac.uk/bialab/GlasContest https://github.com/pkainz/glandsegmentation-models https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/glascontest	LeNet-5 architecture
Sirinukunwattana et al. (2016) [48]	–	A dataset of colorectal adenocarcinoma images, consisting of more than 20,000 annotated nuclei belonging to four different classes	A whole-slide image is first divided into small tiles of size 1,000 × 1,000 pixels	https://www.bioconductor.org/pack-ages/CRImage/	The authors described six layers and used MatConvNet
Komeda et al. (2017) [45]	–	A total of 1,200 images from cases of colonoscopy	256 × 256 pixels	–	An architecture is described

Again, this goal cannot be achieved without the collaboration of medical experts, pathologists, computer programmers, and engineers designing these systems.

Third, future studies should give more attention to the assessment of accuracy, sensitivity of CNNs in evaluating gastrointestinal cancers. Ideally, a study should use two or three different methods and compare the accuracy parameters for the same set of images on using these different methods. Currently, we lack such studies in the literature, and so any comparison of accuracy is not optimum because multiple variables other than the method differ.

7.5 Conclusions

With the increasing research on CNNs and their application in assessing gastrointestinal and liver cancers images, there is a need to carefully evaluate their accuracy and define future research directions. The current studies showed that universities and research institutes in Japan, China, and the United States have pioneered research in this area. Although the current review assessed major cancers of the gastrointestinal system and the liver, the number of studies conducted thus far is small and limited, and more research is needed to answer questions about the accuracy and sensitivity of CNN models. The CNN algorithm indicated the ability to use deep learning in segmentation, classification, and lesion detection in medical images of upper and lower gastrointestinal endoscopies, and radiological and pathological images of common cancers. However, several deficiencies were observed in these studies. In most studies, there was no balance in the content of methods between description of patients involved, the medical components, and the technical computer-related components. Furthermore, there were no studies comparing CNNs with other models, such as MTANNs, regarding the accuracy and sensitivity of each model on the same set of images. Therefore, future studies that focus on these areas of deficiencies and multi-institute, multicenter collaboration with a large number of patients in each of the gastrointestinal cancers should be encouraged. This is particularly important in view of the growing demand of CNNs in medical sciences.

Acknowledgement The author would like to thank Dr. Sarah Azer, St Vincent Hospital, University of Melbourne, for her help during writing this research article.

Funding/Support This work was funded by the College of Medicine Research Center, Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia.

Conflict-of-Interest Statement The author declares that he has no conflict of interest.

References

1. Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S., Abe, O.: Deep learning with convolutional neural network in radiology. Jpn. J. Radiol. **36**(4), 257–272 (2018). https://doi.org/10.1007/s11604-018-0726-3
2. Yu, S., Xiao, D., Kanagasingam, Y.: Exudate detection for diabetic retinopathy with convolutional neural networks. In: Conference Proceedings IEEE Engineering in Medicine and Biology Society, Jul 2017 (pp. 1744–1747). https://doi.org/10.1109/EMBC.2017.8037180
3. Gonzalez, D.I.: DermaKNet: incorporating the knowledge of dermatologists to convolutional neural networks for skinlesion diagnosis. IEEE J. Biomed. Health Inform. (2018). https://doi.org/10.1109/JBHI.2018.2806962
4. Smith, K.P., Kang, A.D., Kirby, J.E.: Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. J. Clin. Microbiol. **56**(3) (2018). https://doi.org/10.1128/JCM.01521-17. pii: e01521–17
5. Blau, N., Klang, E., Kiryati, N., Amitai, M., Portnoy, O., Mayer, A.: Fully automatic detection of renal cysts in abdominal CT scans. Int. J. Comput. Assist. Radiol. Surg. **13**(7), 957–966 (2018). https://doi.org/10.1007/s11548-018-1726-6
6. Ahmad, J., Muhammad, K., Lee, M.Y., Baik, S.W.: Endoscopic image classification and retrieval using clustered convolutional features. Med. Syst. **41**(12), 196 (2017). https://doi.org/10.1007/s10916-017-0836-y
7. Wen, S., Kurc, T.M., Hou, L., Saltz, J.H., Gupta, R.R., Batiste, R., Zhao, T., Nguyen, V., Samaras, D., Zhu, W.: Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images. AMIA Jt. Summits Transl. Sci. Proc. 2018 May 18 (pp. 227–236) (2017). eCollection 2018
8. Li, Y., Shen, L.: Skin lesion analysis towards melanoma detection using deep learning network. Sensors (Basel). **18**(2) (2018). https://doi.org/10.3390/s18020556. pii: E556.
9. Rong, Y., Xiang, D., Zhu, W., Yu, K., Shi, F., Fan, Z., Chen, X.: Surrogate-assisted retinal OCT image classification based on convolutional neural networks. IEEE J. Biomed. Health Inform. (2018). https://doi.org/10.1109/JBHI.2018.2795545
10. Mahmood, F., Durr, N.J.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Ed. Image Anal. **48**, 230–243 (2018). https://doi.org/10.1016/j.media.2018.06.005
11. Zhang, Y., Chandler, D.M., Mou, X.: Quality assessment of screen content images via convolutional-neural-network-based synthetic/natural segmentation. IEEE Trans. Image Process. (2018). https://doi.org/10.1109/TIP.2018.2851390. [Epub ahead of print]
12. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2018. CA Cancer J. Clin. **68**(1), 7–30 (2018). https://doi.org/10.3322/caac.21442
13. National Center for Health Statistics, Centers for Disease Control and Prevention. National Health Interview Surveys, 2000 and 2015. Public Use Data Files 2001. Atlanta, GA: National Center for Health Statistics, Centers for Disease Control and Prevention (2016)
14. Wolf, A.M.D., Fontham, E.T.H., Church, T.R., Flowers, C.R., Guerra, C.E., LaMonte, S.J., Etzioni, R., McKenna, M.T., Oeffinger, K.C., Shih, Y.T., Walter, L.C., Andrews, K.S., Brawley, O.W., Brooks, D., Fedewa, S.A., Manassaram-Baptiste, D., Siegel, R.L., Wender, R.C., Smith, R.A.: Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA Cancer J. Clin. (2018). https://doi.org/10.3322/caac.21457
15. Leufkens, A.M., van Oijen, M.G., Vleggaar, F.P., Siersema, P.D.: Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy **44**(5), 470–475 (2012)
16. Yun, G.Y., Eun, H.S., Kim, J.S., Joo, J.S., Kang, S.H., Moon, H.S., Lee, E.S., Kim, S.H., Sung, J.K., Lee, B.S., Jeong, H.Y.: Colonoscopic withdrawal time and adenoma detection in the right colon. Medicine (Baltimore) **97**(35), e12113 (2018)
17. El-Halabi, M.M., Rex, D.K., Saito, A., Eckert, G.J., Kahi, C.J.: Defining adenoma detection rate benchmarks in average-risk male veterans. Gastrointest. Endosc. (2018). pii: S0016-5107(18)32979-1.

18. Corley, D.A., Jensen, C.D., Marks, A.R., Zhao, W.K., Lee, J.K., Doubeni, C.A., Zauber, A.G., de Boer, J., Fireman, B.H., Schottinger, J.E., Quinn, V.P., Ghai, N.R., Levin, T.R., Quesenberry, C.P.: Adenoma detection rate and risk of colorectal cancer and death. N. Engl. J. Med. **370**(14), 1298–1306 (2014). https://doi.org/10.1056/NEJMoa1309
19. Atkins, L., Hunkeler, E.M., Jensen, C.D., Michie, S., Lee, J.K., Doubeni, C.A., Zauber, A.G., Levin, T.R., Quinn, V.P., Corley, D.A.: Factors influencing variation in physician adenoma detection rates: a theory-based approach for performance improvement. Gastrointest. Endosc. **83**(3), 617–26.e2 (2016)
20. Lee, T.J., Rees, C.J., Blanks, R.G., Moss, S.M., Nickerson, C., Wright, K.C., James, P.W., McNally, R.J., Patnick, J., Rutter, M.D.: Colonoscopic factors associated with adenoma detection in a national colorectal cancer screening program. Endoscopy **46**(3), 203–211 (2014)
21. Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G.J.A., Maciejewski, R., Polkowski, W.P.: Gastric cancer: epidemiology, prevention, classification, and treatment. Cancer Manag. Res. **7**(10), 239–248 (2018). https://doi.org/10.2147/CMAR.S149619
22. O'Donovan, M., Fitzgerald, R.C.: Screening for Barrett's Esophagus: are new high-volume methods feasible? Dig. Dis. Sci. (2018). https://doi.org/10.1007/s10620-018-5192-3
23. Ghouri, Y.A., Mian, I., Rowe, J.H.: Review of hepatocellular carcinoma: epidemiology, etiology, and carcinogenesis. J. Carcinog. **29**(16), 1 (2017). https://doi.org/10.4103/jcar.JCar_9_16.eCollection2017.Review
24. Omata, M., Cheng, A.L., Kokudo, N., Kudo, M., Lee, J.M., Jia, J., Tateishi, R., Han, K.H., Chawla, Y.K., Shiina, S., Jafri, W., Payawal, D.A., Ohki, T., Ogasawara, S., Chen, P.J., Lesmana, C.R.A., Lesmana, L.A., Gani, R.A., Obi, S., Dokmeci, A.K., Sarin, S.K.: Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. Hepatol. Int. **11**(4), 317–370 (2017). https://doi.org/10.1007/s12072-017-9799-9
25. Miao, Z.F., Liu, X.Y., Wang, Z.N., Zhao, T.T., Xu, Y.Y., Song, Y.X., Huang, J.Y., Xu, H., Xu, H.M.: Effect of neoadjuvant chemotherapy in patients with gastric cancer: a PRISMA-compliant systematic review and meta-analysis. BMC Cancer. **18**(1), 118 (2018). https://doi.org/10.1186/s12885-018-4027-0
26. Azer, S.A., Azer, D.: Group interaction in problem-based learning tutorials: a systematic review. Eur. J. Dent. Educ. **19**(4), 194–208 (2015). https://doi.org/10.1111/eje.12121
27. Cohen, J.L., Thomas, J., Paradkar, D., Rotunda, A., Walker, P.S., Beddingfield, F.C., Philip, A., Davis, P.G., Yalamanchili, R.: An interrater and intrarater reliability study of 3 photographic scales for the classification of perioral aesthetic features. Dermatol. Surg. **40**(6), 663–670 (2014). https://doi.org/10.1111/dsu.0000000000000008
28. Takiyama, H., Ozawa, T., Ishihara, S., Fujishiro, M., Shichijo, S., Nomura, S., Miura, M., Tada, T.: Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci. Rep. **8**(1), 7497 (2018). https://doi.org/10.1038/s41598-018-25842-6
29. Fechter, T., Adebahr, S., Baltas, D., Ben Ayed, I., Desrosiers, C., Dolz, J.: Esophagus segmentation in CT via 3D fully convolutional neural network and random walk. Med. Phys. **44**(12), 6341–6352 (2017). https://doi.org/10.1002/mp.12593
30. Xue, D.X., Zhang, R., Feng, H., Wang, Y.L.: CNN-SVM for microvascular morphological type recognition with data augmentation. J. Med. Biol. Eng. **36**(6), 755–764 (2016). https://doi.org/10.1007/s40846-016-0182-4
31. Shichijo, S., Nomura, S., Aoyama, K., Nishikawa, Y., Miura, M., Shinagawa, T., Takiyama, H., Tanimoto, T., Ishihara, S., Matsuo, K., Tada, T.: Application of convolutional neural networks in the diagnosis of helicobacter pylori infection based on endoscopic images. EBioMedicine **25**, 106–111 (2017). https://doi.org/10.1016/j.ebiom.2017.10.014
32. Itoh, T., Kawahira, H., Nakashima, H., Yata, N.: Deep learning analyzes helicobacter pylori infection by upper gastrointestinal endoscopy images. Endosc. Int. Open. **6**(2), E139–E144 (2018). https://doi.org/10.1055/s-0043-120830
33. Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., Tada, T.: Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer **21**(4), 653–660 (2018). https://doi.org/10.1007/s10120-018-0793-2. **Epub 2018 Jan 15**

34. Zhang, X., Hu, W., Chen, F., Liu, J., Yang, Y., Wang, L., Duan, H., Si, J.: Gastric precancerous diseases classification using CNN with a concise model. PLoS One **12**(9), e0185508 (2017). https://doi.org/10.1371/journal.pone.0185508
35. Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q.: Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks. Med. Image Comput. Assist. Interv. **9901**, 442–450 (2016). https://doi.org/10.1007/978-3-319-46723-8_51
36. Xing, F., Xie, Y., Yang, L.: An automatic learning-based framework for robust nucleus segmentation. IEEE **35**(2), 550–566 (2016). https://doi.org/10.1109/TMI.2015.2481436
37. Liu, X., Guo, S., Yang, B., Ma, S., Zhang, H., Li, J., Sun, C., Jin, L., Li, X., Yang, Q., Fu, Y.: Automatic organ segmentation for ct scans based on super-pixel and convolutional neural networks. J. Digit. Imaging (2018). https://doi.org/10.1007/s10278-018-0052-4
38. Liu X, Song JL, Wang SH, Zhao JW, Chen YQ. Learning to Diagnose Cirrhosis with Liver Capsule Guided Ultrasound Image Classification. Sensors (Basel). 2017 Jan 13;17(1). pii: E149. https://doi.org/10.3390/s17010149
39. Ibragimov, B., Toesca, D., Chang, D., Koong, A., Xing, L.: Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. Phys. Med. Biol. **62**(23), 8943–8958 (2017). https://doi.org/10.1088/1361-6560/aa9262
40. Yasaka, K., Akai, H., Abe, O., Kiryu, S.: Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology **286**(3), 887–896 (2018). https://doi.org/10.1148/radiol.2017170706
41. Qin, W., Wu, J., Han, F., Yuan, Y., Zhao, W., Ibragimov, B., Gu, J., Xing, L.: Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. Phys. Med. Biol. **63**(9), 095017 (2018). https://doi.org/10.1088/1361-6560/aabd19
42. Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P.: Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology (2018). pii: S0016-5085(18)34659-6. https://doi.org/10.1053/j.gastro.2018.06.037
43. Billah, M., Waheed, S., Rahman, M.M.: An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. Int. J. Biomed. Imaging **2017**, 9545920 (2017)
44. Zhang, R., Zheng, Y., Mark, T.W.C., Yu, R., Wong, S.H., Lau, J.Y.W., Poon, C.C.Y.: Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE **21**(1), 41–47 (2017). https://doi.org/10.1109/JBHI.2016.2635662
45. Komeda, Y., Handa, H., Watanabe, T., Nomura, T., Kitahashi, M., Sakurai, T., Okamoto, A., Minami, T., Kono, M., Arizumi, T., Takenaka, M., Hagiwara, S., Matsui, S., Nishida, N., Kashida, H., Kudo, M.: Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology **93**(Suppl 1), 30–34 (2017). https://doi.org/10.1159/000481227
46. Haj-Hassan, H., Chaddad, A., Harkouss, Y., Desrosiers, C., Toews, M., Tanougast, C.: Classifications of multispectral colorectal cancer tissues using convolution neural network. J. Pathol. Inform. **28**(8), 1 (2017). https://doi.org/10.4103/jpi.jpi_47_16
47. Kainz, P., Pfeiffer, M., Urschler, M.: Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. PeerJ. **3**(5), e3874 (2017). https://doi.org/10.7717/peerj.3874
48. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE **15**(5), 1196–1206 (2016). https://doi.org/10.1109/TMI.2016.2525803
49. Men, K., Dai, J., Li, Y.: Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med. Phys. **44**(12), 6377–6388 (2017)
50. Suzuki, K.: Overview of deep learning in medical imaging. Radiol. Phys. Technol. **10**(3), 257–273 (2017). https://doi.org/10.1007/s12194-017-0406-5
51. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights Imaging (2018). https://doi.org/10.1007/s13244-018-0639-9

52. Azer, S.A., Dupras, D.M., Azer, S.: Writing for publication in medical education in high impact journals. Eur. Rev. Med. Pharmacol. Sci. **18**(19), 2966–2981 (2014)
53. Suzuki, K., Yoshida, H., Näppi, J., Dachman, A.H.: Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: suppression of rectal tubes. Med. Phys. **33**(10), 3814–3824 (2006)
54. Wang, H., Zhao, T., Li, L.C., Pan, H., Liu, W., Gao, H., Han, F., Wang, Y., Qi, Y., Liang, Z.: A hybrid CNN feature model for pulmonary nodule malignancy risk differentiation. J. Xray Sci. Technol. 26(2), 171–187 (2018). https://doi.org/10.3233/XST-17302
55. Nanni, L., Ghidoni, S., Brahnam, S.: Ensemble of convolutional neural networks for bioimage classification. Appl. Comput. Inform. (2018). https://doi.org/10.1016/j.aci.2018.06.002
56. Dawson, D.V., Pihlstrom, B.L., Blanchette, D.R.: Understanding and evaluating meta-analysis. J. Am. Dent. Assoc. **147**(4), 264–270 (2016). https://doi.org/10.1016/j.adaj.2015.10.023
57. Chong, S.W., Collins, N.F., Wu, C.Y., Liskaser, G.M., Peyton, P.J.: The relationship between study findings and publication outcome in anesthesia research: a retrospective observational study examining publication bias. Can. J. Anaesth. **63**(6), 682–690 (2016). https://doi.org/10.1007/s12630-016-0631-0
58. Wagner, C.S., Park, H.W., Leydesdorff, L.: The continuing growth of global cooperation networks in research: a conundrum for national governments. PLoS One **10**(7), e0131816 (2015). https://doi.org/10.1371/journal.pone.0131816

Chapter 8
Supervised CNN Strategies for Optical Image Segmentation and Classification in Interventional Medicine

Sara Moccia, Luca Romeo, Lucia Migliorelli, Emanuele Frontoni and Primo Zingaretti

Abstract The analysis of interventional images is a topic of high interest for the medical-image analysis community. Such an analysis may provide interventional-medicine professionals with both decision support and context awareness, with the final goal of improving patient safety. The aim of this chapter is to give an overview of some of the most recent approaches (up to 2018) in the field, with a focus on Convolutional Neural Networks (CNNs) for both segmentation and classification tasks. For each approach, summary tables are presented reporting the used dataset, involved anatomical region and achieved performance. Benefits and disadvantages of each approach are highlighted and discussed. Available datasets for algorithm training and testing and commonly used performance metrics are summarized to offer a source of information for researchers that are approaching the field of interventional-image analysis. The advancements in deep learning for medical-image analysis are

S. Moccia (✉) · L. Romeo · L. Migliorelli · E. Frontoni · P. Zingaretti
Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
e-mail: s.moccia@univpm.it

L. Romeo
e-mail: l.romeo@univpm.it

L. Migliorelli
e-mail: l.migliorelli@pm.univpm.it

E. Frontoni
e-mail: e.frontoni@univpm.it

P. Zingaretti
e-mail: zinga@dii.univpm.it

S. Moccia
Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy

L. Romeo
Department of Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa, Italy

Department of Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa, Italy

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_8

involving more and more the interventional-medicine field. However, these advancements are undeniably slower than in other fields (e.g. preoperative-image analysis) and considerable work still needs to be done in order to provide clinicians with all possible support during interventional-medicine procedures.

8.1 Introduction to Optical-Image Analysis in Interventional Medicine

Nowadays, the surgeon's decision process combines (i) pre-operative qualitative analysis of patient-specific anatomy and physiology, retrieved from imaging systems and sensors, and (ii) surgeon's prior knowledge about medical rules and statistics [1]. Such information is used to build an implicit patient's model and define a surgical plan. After-surgery, surgical outcomes are qualitatively evaluated and statistically analyzed to improve treatment effectiveness and eventually change treatment protocol (Fig. 8.1).

Advancements in intra-operative imaging systems and computer-based analysis allowed to acquire more and more information on patient's anatomy and physiology to eventually update the surgical plan directly in the operating room (OR). In fact, surgeons commonly exploit optical imaging when performing interventional-

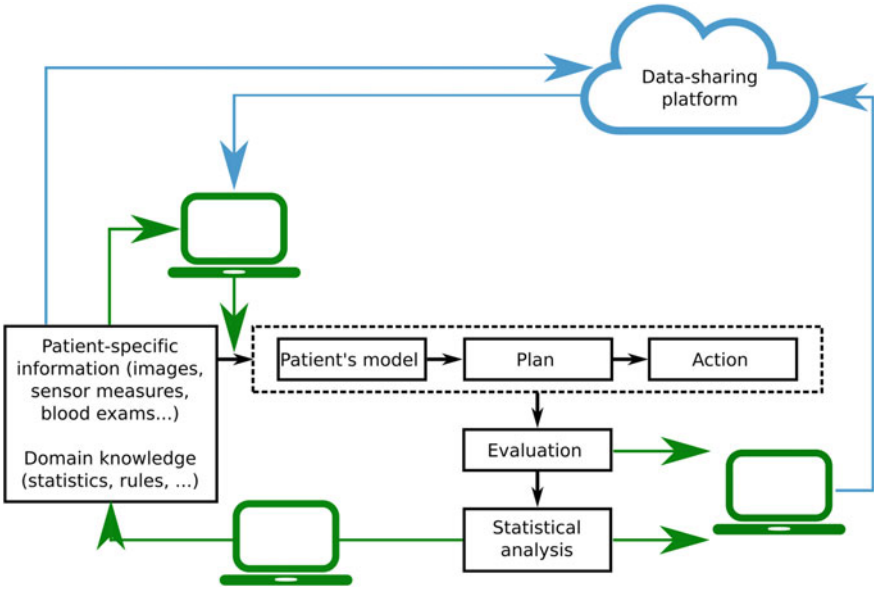


Fig. 8.1 Surgical data science integration in the interventional-medicine workflow allows objective decision-making and quantitative evaluation of the surgical outcomes

medicine procedures for obtaining both diagnostic support and context awareness in a non-invasive way [2]. New imaging devices that combine advanced sensors and increased computational power are constantly introduced in the OR, e.g., multispectral [3], narrow-band [4], and spectroscopy imaging [5]. Endoscopic cameras today allow to perform minimally invasive surgery (MIS) improving post-operative patient's prognosis and quality of life [1]. Robotic MIS is gradually emerging as a powerful solution to further improve treatment quality, and is already the state of the art in specific fields (e.g., urology) [6].

As a natural result of the massive introduction of imaging devices in the OR, an almost unlimited amount of electronic patient records are available [1]. These data can be processed in a quantitative way to further increase safety, effectiveness and efficiency of surgical care [2]. Moreover, as observed in [7], the Internet-Of-Things revolution has the healthcare domain as one of the most promising field, with infinite opportunities arising from data sharing among hospitals, care-givers and patients. Indeed, data sharing can provide the surgeons with statistics from other patients shared among care centers and this information can integrate the patient-specific (local) data.

A primary goal of the medical image analysis community is to organize, analyze and model such huge amount of data to enhance the quality of interventional healthcare [2]. In this context, surgical data science (SDS) aims at supporting health specialists through a quantitative processing of intra-operative images to implement

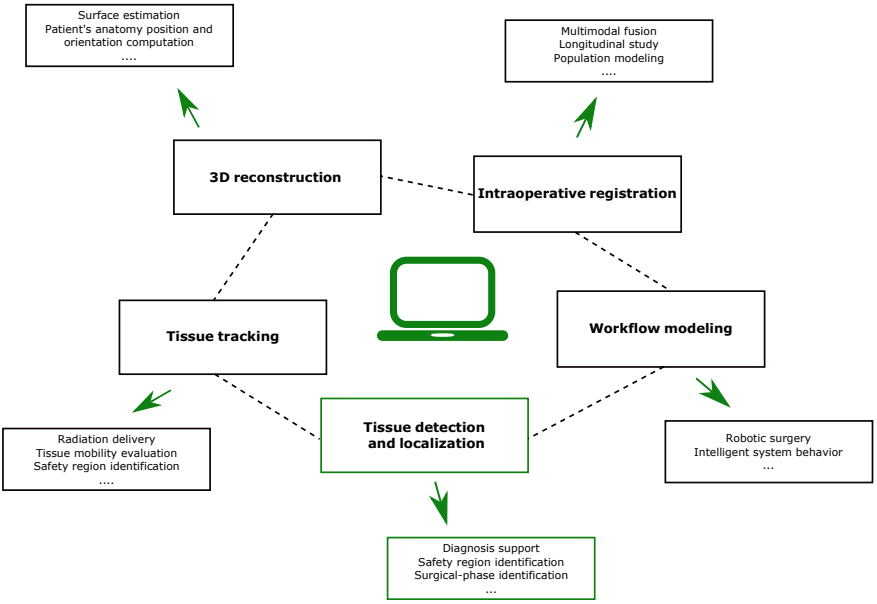


Fig. 8.2 Some of the major opportunities that surgical data science offers to interventional medicine. Blocks highlighted in green identify the main topics of this paper

(Fig. 8.2): tissue tracking [8], 3D reconstruction [9], intra-operative registration [10], workflow modeling [11], detection and localization of anatomical structures [12] or/and surgical instrumentation [13].

In addition to challenges related to intra- and inter-patient variability in biological tissues (especially in presence of pathologies), the processing of optical images acquired during interventional medicine presents further challenges, such as high sensor noise, varying illumination levels, organ movement, different pose of the acquisition sensor with respect to the tissues and presence of blood, smoke and surgical tools in the field of view.

To tackle the high variability of intra-operative optical images, SDS methods and principles heavily build on machine learning (ML) [2]. The medical domain-specific knowledge can be encoded in a ML-based model through a learning process based on the description of cases solved in the past. The model can:

- Offer decision support [11], e.g., by assisting the clinician when diagnosing new patients to improve the diagnostic speed, accuracy and/or reliability;
- Provide context awareness [14], e.g., for autonomous assistance and collaborative robots in MIS to improve safety, quality and efficiency of care.

More recently, deep learning (DL) approaches based on Convolutional Neural Networks (CNNs) for the analysis of interventional-medicine images drew the attention of the SDS community. Remarkable results were obtained in skin-cancer classification [15], polyp detection [16], retinal image analysis [17], and vessel segmentation [18], where large and labeled datasets are publicly available for DL model training. With respect to standard ML approaches to medical optical-image analysis, which require to extract high-level complex features (Sect. 8.1.2), CNNs tackle the classification and segmentation problems from a different point of view and represent the image as a nested hierarchy of simpler features that are automatically learned from the images during the training phase (Fig. 8.3).

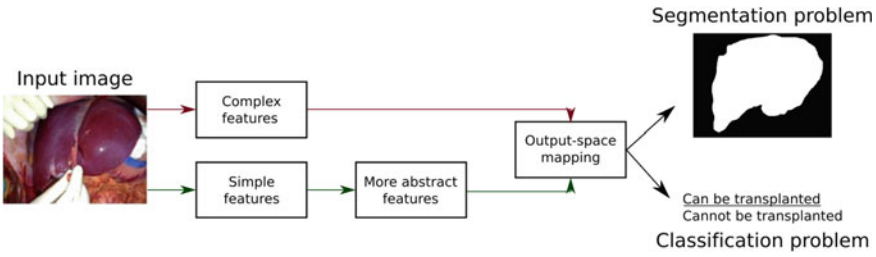


Fig. 8.3 Image segmentation and classification workflows for standard machine-learning (red arrows) and deep-learning (green arrows) approaches

8.1.1 Aim of the Survey

As the use of CNNs in the field of intra-operative optical image analysis is rapidly growing, the primary goal of this review is to provide an up-to-date source of information about its current state in the literature, with a specific focus on tissue classification and segmentation approaches for decision support and context awareness during interventional-medicine procedures.

Reviews in the field of DL for medical image analysis have been previously proposed, but mainly for applications related to anatomical images (such as computed-tomography or magnetic-resonance images) [19], while very few to describe the specific state of the art related to optical images acquired during interventional-medicine procedures. These latter ones only focus of specific anatomical regions without giving an integral vision of the challenges and advancements related to intra-operative tissue analysis. Examples include [20], that surveys methods for gastrointestinal-image analysis from a clinical point of view (more than from a methodological one). In [21, 22], algorithms for polyp and Barrett's esophagus detection are discussed, respectively, focusing on model-based and standard ML algorithms, leaving few space for DL strategies.

This survey may represent a salient resource for researchers in the field of SDS who wants to face up to the problem of intra-operative tissue analysis with DL. It analyzes almost fifty articles published from 2015, both from the methodological and application point of view. After a short introduction (Sect. 8.1.2) on last-decade methodologies, which mostly dealt with standard ML approaches, a short overview on CNNs is given (Sect. 8.1.3).

Considering the importance of having a proper and large training set to encode image and tissue variability when performing tissue classification and segmentation, a section to list and analyze the publicly available and labeled datasets is also included, along with a list of the most common metrics to evaluate algorithm performance in a fair and consistent way (Sect. 8.1.4).

CNN-based methodologies to image analysis are grouped in two categories: image segmentation (Sect. 8.2) and image classification (Sect. 8.3). As the majority of datasets built for interventional-medicine segmentation include also surgical-tool annotation, surgical-tool segmentation strategies are included in Sect. 8.2, too. In each category, articles are further split according to their clinical tasks. Finally, Sect. 8.4 concludes this paper summarizing the main findings and presenting open challenges and future research direction.

8.1.2 Previous Approaches to Tissue Segmentation and Classification

During the last decades, standard ML models for tissue classification typically applied (i) automated image analysis to extract a vector of quantitative, hand-designed, fea-

tures to characterize the relevant image content and (ii) a pattern classifier to map the features to the output space to determine the category to which the extracted feature vector belongs, e.g., malignant/healthy tissue.

The most exploited features were built from intensity, textural and derivative-based information [23]. Intensity-based features aimed at encoding information related to the prevalent intensity components in the image and were mainly based on intensity histogram, mean, variance and entropy. These features were commonly combined with textural features, which encoded tissue appearance, structure and arrangement [24]. Textural features included local binary patterns [25], gray-level co-occurrence matrices [26] and histograms of oriented gradients [27]. Other popular features were obtained with filtering-based approaches, such as matched filtering and wavelet analysis,which have been widely used for polyp classification [28]. Similarly, derivative-based approaches built derivative-filters to extract image spatial derivatives, such as gradient and Laplacian, e.g., to highlight tissue edges [29].

As for pattern classifiers, several solutions were exploited. First attempts were based on probabilistic approaches (i.e., Naive Bayes) [30]. Similarly, perceptron-based algorithms have been widely used, e.g. for polyp detection in endoscopic images [31]. Tree-based algorithms and kernel based methods (i.e., support vector machine) were probably among the most widely used classifiers. These algorithms showed promising performance for tissue classification in several fields (e.g., abdominal-tissue segmentation and classification [24, 26, 32].

8.1.3 Background on Convolutional Neural Networks (CNNs)

As in traditional neural networks, a CNN is a sequence of layers, where the convolutional one is the most peculiar. As pointed out in [33], convolution leverages three important ideas that can help improving classification and segmentation tasks with respect to traditional ML approaches (based on neural networks):

1. *Sparse interactions*. While for traditional networks every output unit interacts with every input unit, CNNs typically have sparse connections.
2. *Parameter sharing*. Rather than learning a separate set of features for every image location, only one set is learned, reasonably assuming that it is independent from the image location.
3. *Equivariant representations*. From the parameter-sharing property, the convolution equivariance to translation arises (i.e., if the input changes, the output changes in the same way).

Using convolutional layers results in fewer parameters to store and thus in reduced memory consumption, higher statistical efficiency and fewer operations to accomplish for output prediction.

In addition to convolution, CNNs commonly implement pooling between successive convolutional layers. With pooling, the output of the net at a certain location is

replaced by a summary statistic of its nearby outputs (e.g., maximum value in case of max pooling). Implementing pooling is equivalent to perform downscaling, thus allowing noise smoothing and making the CNN invariant to small translations of the input.

Regarding image segmentation, today the most successful solutions exploit fully-convolutional neural networks (FCNNs), which allow a faster and more accurate segmentation. FCNNs were first presented in [34] and up to now several architectures, such as UNet [35], SegNet [36] and modified version of ResNet [37], showed remarkable segmentation performance.

For classification tasks, CNNs usually end with one or more fully-connected (dense) layers, i.e., layers where all the units have connection with the units of the previous layer (as in standard neural networks). The number of output units for the last layer coincides with the number of classes (e.g., two units for a binary classification problem such as healthy vs pathological tissue). From the first CNN model for image classification (i.e., LeNet5 [38]), today milestone architectures are Alexnet [39], GoogleNet [40], VGG16 [41] and, more recently, fractal CNNs [42] and residual CNNs such as ResNet [43].

CNN based models were proposed for natural-image analysis, probably because of the availability of huge annotated datasets such as Imagenet ([1]). To take full advantage of the trained models (i.e., CNN weights) available online, a common strategy in interventional medicine imaging analysis is to implement fine tuning. Fine tuning consists in adapting the CNN weights learned with huge natural-image datasets by re-training the last CNN layers with the medical image dataset [15].

8.1.4 Available Datasets and Performance Metrics

Considering the potentiality of learning algorithms to tackle the intra-operative image variability, collecting large quantity of annotated datasets for algorithm training became crucial. Indeed, several international organizations constantly work to collect and label, in a consistent manner, high-quality data recorded during interventional-medicine procedures. However, this positive trend still concerns only few anatomical regions (Table 8.1).

In parallel to the manual annotation of medical datasets, the SDS community is also studying how crowd- powered algorithm collaboration could be used to annotate large-scale medical images, as to moderate the surgeon involvement in the time-consuming annotation process [44].

Segmentation and classification performance is commonly evaluated with respect to the manual annotation performed by expert clinicians. To attenuate intra-subject variability when performing the manual annotation, a combination of annotation by multiple experts is usually employed [45]. When evaluating the algorithm performance with respect to manual annotation, a contingency table with true positive

[1] www.image-net.org/.

Table 8.1 List of available datasets

Name	Classification and segmentation task	Link
ISBI 2016, 2017	Celiac disease	https://aidasub-cleceliachy.grand-challenge.org/home/
	Gastric cancer	https://aidasub-chromogastro.grand-challenge.org/
	Barrett's esophagus	https://aidasub-clebarrett.grand-challenge.org/home/
KID	Gastrointestinal lesions	https://mdss.uth.gr/datasets/endoscopy/kid/
CVC colon DB	Colon polyps	http://mv.cvc.uab.es/projects/colon-qa/cvccolondb
EndoScene	Colon polyps	https://github.com/jbernoz/deeppolyp
MICCAI EndoVis	Colon polyps (ASU-Mayo)	https://endovis.grand-challenge.org
	Vascular and inflammatory gastrointestinal lesions (GIANA)	
	Outer edge of kidney (Kidney Boundary Detection)	
	Barrett's esophagus (Early Barrett's cancer detection)	
Cervix dataset	Cervix type	https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data
ISIC 2016, 2017, 2018	Dermoscopic images	https://www.isic-archive.com

(TP), true negative (TN), false negative (FN) and false positive (FP) is commonly used. The positive and negative samples refer, in turn, to pixels within and outside the segmented region (segmentation task) or images belonging to diseased and healthy class (classification task) according to the manual annotation. Commonly exploited metrics that are computed from the contingency table are accuracy (Acc), sensitivity (Se), specificity (Sp) and precision (Pr):

$$Acc = \frac{TP + TN}{n} \tag{8.1}$$

$$Se = \frac{TP}{TP + FN} \tag{8.2}$$

$$Sp = \frac{TN}{TN + FP} \tag{8.3}$$

$$Pr = \frac{TP}{TP + FP} \tag{8.4}$$

being n the total number of pixels (segmentation task) or images (classification task).

The area under (AU) the Receiver Operating Characteristic (ROC) is also used as a metric (especially with skewed classes), where the ROC describes the performance of a binary classifier system as its discrimination threshold is varied.

When dealing with segmentation, further measures based on spatial overlapping can be used, too. The most used ones are the Dice Similarity Coefficient (DSC), also known as $F1\_score$, and the Jaccard coefficient (JC):

$$DSC = \frac{2TP}{FP + FN + 2TP} \tag{8.5}$$

$$JC = \frac{DSC}{2 - DSC} \tag{8.6}$$

8.2 Optical-Image Segmentation

This section will survey approaches for the segmentation of images acquired during interventional-medicine procedures. For each segmentation approach, Table 8.2 lists the relative anatomical region, image dataset, segmentation task and performance metrics. Figure 8.4 shows visual samples for skin, polyp and surgical-tool analysis.

Skin lesions

Following the first CNN-based approach to pathological skin-image analysis, mainly dealing with classification tasks [15], several methods for lesion segmentation have been proposed. In [47], an encoder-decoder network is proposed to melanoma segmentation. The network is based on U-Net and includes skip connections, as in

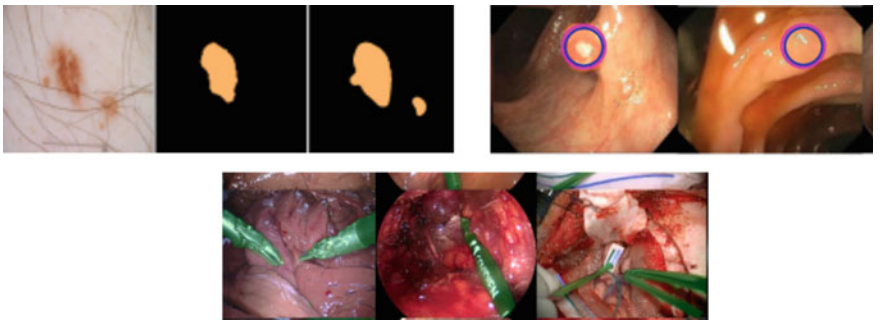


Fig. 8.4 Segmentation samples for skin, polyp and surgical instruments. Images adapted from [47, 55, 57]

Table 8.2 Summary table for image-segmentation approaches

Article	Anatomical region	Dataset	Task	Performance metrics					
				Acc	*Se*	*Sp*	*Pr*	*DSC*	*JC*
Bi et al. (2017) [46]	Skin	ISIC 2017	Lesions						0.794
Sarker et al. (2018) [47]	Skin	ISBI 2016	Lesions	0.984	0.945	0.992		0.955	0.913
	Skin	ISBI 2017	Lesions	0.936	0.816	0.983		0.878	0.782
Mirikharaji et al. (2018) [48]	Skin	ISBI 2017	Lesions	0.938	0.855	0.973		0.857	0.773
Ghosh et al. (2018) [49]	Colon	KID	Lesions	0.944					
Wickstrom et al. (2018) [50]	Colon	EndoScene	Polyps	0.949					
Vazquez et al. (2016) [51]	Colon	EndoScene	Polyps	0.930					
Brandao et al. (2018) [52]	Colon	MICCAI EndoVis and CVC colon DB	Polyps						
Laina et al. (2017) [53]	Gastrointestinal tract	MICCAI EndoVis	Surgical tool	0.926	0.862	0.99		0.889	
Attia et al. (2017) [54]	Gastrointestinal tract	MICCAI EndoVis	Surgical tool	0.933					0.827
Garcia et al. (2015) [55]	Gastrointestinal tract	MICCAI EndoVis (non real time)	Surgical tool	0.837	0.722	0.952			
	Gastrointestinal tract	MICCAI EndoVis (real time)	Surgical tool	0.883	0.878	0.887			
Milletari et al. (2018) [56]	Gastrointestinal tract	MICCAI EndoVis	Surgical tool	0.978	0.888	0.988		0.895	

Acc accuracy, *Se* sensitivity, *Sp* specificity, *Pr* precision, *DSC* Dice similarity coefficient, *JC* Jaccard coefficient

ResNets, and dilated convolution [58]. ResNet is also used in [46]. A similar approach is proposed in [48], with the main innovation of including shape priors in the loss function used to train the FCNN. This yields to faster convergence and more accurate segmentation results. U-Net is also exploited in [59], where a nested architecture is proposed by optimizing a loss function that allows handling partial image labeling in confocal microscopy skin images.

Gastrointestinal lesions al polyps

A benchmark analysis for FCNN-based polyp segmentation is proposed in [51], using one of the first FCNN model in the literature [34]. In [49, 50], a modified version of SegNet is proposed for pixel-wise polyp and bleeding segmentation in wireless-endoscopy images, respectively. Polyp detection is achieved with SegNet in [49], too. In [50] a similar approach is investigated for polyp detection, with further segmentation-uncertainty estimation via Monte Carlo dropout and model interpretability analysis by highlighting descriptive regions in the input images with guided backpropagation [60].

Two parallel custom-built CNNs (for edge detection and lesion classification) are described in [61] to allow Hookworm disease detection in wireless endoscopic images. In [57], temporal information is included in the polyp detection process by building a 3D CNN. Experimental results show an improvement in the detection performance with respect to approaches based on single-frame processing.

Depth information is exploited in [52] as an additional input channel to FCNN architectures based on VGG16 and Resnet to the RGB information, experimentally demonstrating improved performance. Growing interest in the field is also reserved to automatic depth prediction with CNNs for 3D colon-shape reconstruction [62–64].

Surgical tools for gastrointestinal surgery

One of the first real-time FCNN-based approaches to the segmentation of non-rigid surgical tools was proposed in [55], where SegNet was adapted and fine-tuned to segment surgical tool in endoscopic images. A similar approach is proposed in [53], where the FCNN encoder is inspired by ResNet, and the decoder one has two branches for generating both the instrument segmentation mask and its articulated 2D pose.

In [86], a U-Net based architecture to surgical tool segmentation is proposed. The FCNN is modified to allow multiple instrument segmentation. The FCNN is in series with a second regressor network to regress the instrument pose.

Recurrent networks are used in [54, 56], where an encoder-decoder FCNN inspired to U-Net is combined with Long Short Term Memory (LSTM) to provide instrument segmentation in endoscopic images while encoding temporal dependencies. This methodology results in higher accuracy than approaches based on non-recurrent networks. With the same aim, CNNs with 3D kernels have been proposed in [87] for instrument pose estimations.

8.3 Optical-Image Classification

This section will survey approaches for the classification of images acquired during interventional-medicine procedures. For each segmentation approach, Table 8.3 lists the relative anatomical region, image dataset, classification task and performance metrics. Figure 8.5 shows visual samples for skin, gastrointestinal and oral-cavity lesion classification.

First approaches to CNN-based tissue classification exploited CNN simply to extract learned features, which then will be used for tissue classification with standard ML-approaches introduced in Sect. 8.1.2 [19]. This was mainly related to the small numerosity of image datasets. When larger datasets started to become publicly available, more advanced approaches were investigated, which we will survey hereafter. Accordingly, CNN-based approaches started to be exploited in order to (i) learn discriminative nonlinear features and (ii) classify the optical-images according to such features.

Skin lesions

The work presented in [15] is one of the first approaches to skin-lesion segmentation with CNN, where Google Inception v3 is fine-tuned to detect tumoral skin lesions. A similar approach, which uses VGG16 as classification network, is presented in [66], while ResNet is fine-tuned to classify skin lesions in [46]. A three-branch CNN is proposed in [68] for coarse classification of psoriatic-plaque macro classes. After a common VGG16-like architecture, each branch is deputy to a finer classification of plaque grade. Webly supervised learning is investigated in [67] to deal with data imbalance. An innovative approach to synergic DL has been recently proposed in [65], overcoming state of the art approaches.

Gastrointestinal lesions and polyps

Ulcer and bleeding in wireless endoscopic images are classified in [74, 76] using a sixteen- and ten-layer CNN, respectively. A similar approach is exploited in [77], where the CNN is fed with both endoscopic frames and image priors (Hessian and Laplacian) to improve the classification performance. For the same task, AlexNet is used in [75]. Interesting approaches to weakly-supervised CNN for detection of inflammatory gastrointestinal lesions are proposed in [70, 78], to overcome the problem of limited number of annotated images.

A simple CNN with six stages is used in [71] to classify Barrett's esophagus and neoplasia in endomicroscopy images. In [69] a further advancement is done and Barrett's esophagus frames are classified by fine-tuning ResNet.

One of the first approaches to polyp classification using an end-to-end trained CNN is proposed in [72], where transfer learning is applied to several CNN models, such as VGG16 and Alexnet, outperforming conventional ML methods. An innovative approach to polyp classification is proposed in [73] where a 10-stages CNN architecture that consists of alternated convolutional and dense layers is built and regularized to be rotation-invariant.

Table 8.3 Summary table for image-classification approaches

Article	Anatomical region	Dataset	Task	n of classes	Performance metrics					
					Acc	*Se*	*Sp*	*Pr*	*AUC*	*DSC*
Esteva et al. (2017) [15]	Skin	129450 images	Lesions	2					0.953	
Bi et al. (2017) [46]	Skin	ISIC 2017	Lesions	2					0.855	
	Skin	ISIC 2017	Lesions	2+3 (ensemble)					0.976	
Zhang et al. (2018) [65]	Skin	ISIC 2016	Lesions	2	0.858			0.664	0.818	
Lopez et al. (2017) [66]	Skin	ISIC	Lesions	2		0.787		0.797		
Navarro et al. (2018) [67]	Skin	1300 images	Lesions	10					0.9695	
Pal et al. (2016) [68]	Skin	707 images	Lesions	3	0.589					
Mendel et al. (2017) [69]	Gastrointestinal tract	MICCAI EndoVis 2015	Lesions	2		0.940	0.880			
Georgakopoulos et al. (2016) [70]	Gastrointestinal tract	KID	Lesions	2	0.902	0.926	0.889			
Hong et al. (2017) [71]	Gastrointestinal tract	ISBI 2016	Lesions	3	0.808					
Ribeiro et al. (2016) [72]	Gastrointestinal tract	818 images	Polyps	2	0.936					
Yuan et al. (2018) [73]	Gastrointestinal tract	3000 WCE images	Polyps	2	0.956	0.950		0.963		0.956
Aoki et al. (2018) [74]	Gastrointestinal tract	15800 images	Lesions	2	0.908	0.882	0.909		0.958	
Fan et al. (2018) [75]	Gastrointestinal tract	WCE images	Lesions	2	0.953		0.954		0.909	

(continued)

Table 8.3 (continued)

Article	Anatomical region	Dataset	Task	n of classes	Performance metrics					
					Acc	*Se*	*Sp*	*Pr*	*AUC*	*DSC*
Sekuboyina et al. (2017) [76]	Gastrointestinal tract	137 images	Lesions	8					0.8	
Segui et al. (2016) [77]	Gastrointestinal tract	120K WCE images	Lesions	6	0.96					
Vasilakakis et al. (2018) [78]	Gastrointestinal tract	KID	Lesions	2	0.9					
Itoh et al. (2018) [79]	Gastrointestinal tract	170 images	Other applications (helicobapter pilori)	2		0.867	0.867		0.956	
Yu et al. (2015) [80]	Gastrointestinal tract	1mln real WCE images	Other applications (digestive organs)	3	0.973					
Chen et al. (2017) [81]	Gastrointestinal tract	630k images	Other applications (digestive organs)	3	0.897	0.943	0.900			
Aubreville et al. (2017) [82]	Oral cavity	7894 images	Other applications (cancerous lesions)	2	0.883	0.866	0.900		0.960	
Xu et al. (2016) [83]	Cervix	690 images	Other applications (cervix dysplasia)	2	0.889	0.878	0.900		0.940	
Zou et al. (2015) [84]	Gastrointestinal tract	1 mln real WCE images	Other applications (organ classification)	3	0.955					
Zhou et al. (2017) [85]	Gastrointestinal tract	8800 images	Other applications (celiac disease)	2		1	1			

WCE wireless capsule endoscopy, *Acc* accuracy, *Se* sensitivity, *Sp* specificity, *Pr* precision, *AUC* area under the receiver operating characteristic, *DSC* Dice similarity coefficient

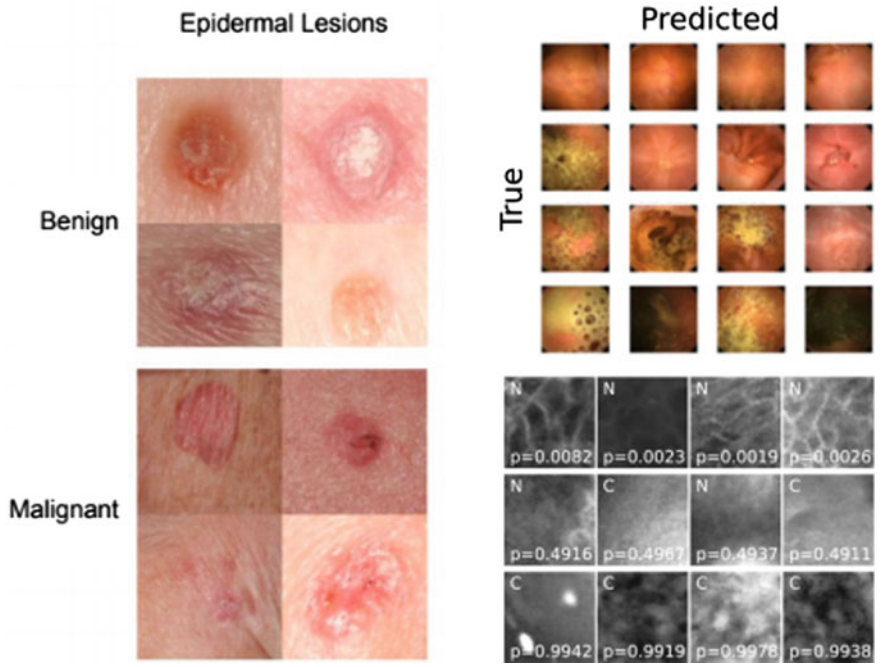


Fig. 8.5 Classification samples for skin lesions, polyp and oral-cavity cancer. Images adapted from [15, 77, 88]

Other applications

CNNs inspired by AlexNet are used in [81, 84] to identify digestive organs. A seven-stage CNN is used in [80] to automatically extract image features that are then classified with extreme ML. Such approach experimentally shows better performance than using a fully-connecting layers, probably due to the small depth of the CNN.

GoogleNet is used in [79], to classify Helicobacter pylori infection in upper gastrointestinal endoscopy images. Fine-tuning is implemented to transfer the recognition capabilities of the GoogleNet to the endoscopic images. A similar approach is used in [85] for celiac disease classification by video endoscopy.

A seven-stages CNN is tested in [82] to classify cancerous tissue in laserendomicroscopy images of the oral cavity, showing improved performance with respect to standard ML-based approaches in the field.

A multimodal CNN-based approach to cervical dysplasia classification is proposed in [83]: this combines both automatic feature extraction with CNNs and data from clinical records.

8.4 Discussion

The efforts in the field of DL applied to optical-image analysis are promising and encouraging, however, several methodological and technical challenges are still open, hampering the translation of these developed methodologies into the actual clinical practice.

From the methodological point of view, it emerged that a comparison of the proposed methodologies is not trivial. There is not a consensus yet on the exploited datasets and the reported performance metrics, which are not consistent among different research articles (see Tables 8.2 and 8.3). Moreover, despite the efforts invested in the analysis of interventional-medicine images, the number of research articles in this field is still lower than that relate to anatomical-image analysis [19].

Regarding the technical challenges, there are several aspects that can be tackled to potentially achieve the goal of robust and reliable tissue classification.

The first aspect deals with hardware design. Indeed, the imaging field is constantly evolving thanks to new optical imaging technologies, such as narrow-band imaging [89] and multispectral imaging [90]. These technologies potentially allow high-quality optical imaging (e.g., in terms of image noise and tissue-background contrast) and have already found interesting applications in the remote-sensing field [91]. However, the use of these technologies is still underrepresented in the medical field with few examples (e.g. [32, 92]).

A second aspect is related to the identification of images to be processed. High noise level in the image, camera movements, tissue deformation and illumination drop lower image quality and make the classification challenging also for the human eye. Similarly, classification algorithms are prone to error when processing uninformative and noisy frames. Solutions have been proposed in the literature, nonetheless they are still limited to few anatomical regions and have to be further investigated [88, 93].

A third point concerns the estimation of the level of classification confidence while increasing the model interpretability, with a view to improve generalization performance. In particular, it has been reported that allowing a system to produce "unknown" results can potentially reduce the number of incorrectly classified cases [94]. In this context, advancements in DL aim to discover patterns sometimes unsighted by physicians [95] while estimating the posterior probability of the prediction. Interestingly, understanding why and how the outcome prediction is made may also help the physician to discover salient predictors involved in the diagnostic process (pattern localization) [96]. However, the introduction of confidence estimation in the medical imaging field has been only marginally explored [32]. DL-model interpretability is still an open research topic and recent approaches aim to increase it by (i) employing sparse CNN models with different loss or penalty functions [97] and (ii) exploiting visual-attention models to predict human eye fixations on images [98].

More generally, as SDS/DL strongly rely upon labeled data, the last aspect is related to the availability of labeled datasets. Indeed, the larger the training dataset, the bigger the chances the classification algorithm will be accurate in classifying

unseen data. While the development of tissue-classification algorithms is strongly advancing in some specific fields (e.g., vascular district [18], and gastrointestinal tract [22]), there are other fields that are incredibly underrepresented in the literature. The most probable reason for this is indeed the lack of large and available labeled databases for algorithm training.

Despite international organizations are active in collecting high-quality annotated datasets, several anatomical districts are still underrepresented, thus limiting the applicability of supervised CNN-based approaches. However, only a fraction of patient-related data is digitized and stored in a structured and standardized way, and data quality assessment is rarely performed [2]. This is probably the main reason why SDS only recently emerged as an active field of research. While shared databases are available for other research fields for advancing research (e.g., the ImageNet dataset, www.image-net.org/), annotated datasets for the SDS community are still limited in number. This can be attributed to regulatory and sociological factors (e.g. data protection and privacy issues) [99]. A second factor deals with medical data annotation, which is typically an expensive process in terms of resources and time [100]. In the last years, several efforts have been made by the SDS community to support research-data sharing and develop crowd- powered algorithms for large-medical-dataset tagging [101, 102]. With a focus on imaging data, data sharing is especially supported by international organizations, such as the MICCAI society, the IEEE Signal Processing Society and the IEEE Engineering in Medicine and Biology Society, which yearly organize *Grand Challenges*[2] and release annotated dataset for algorithm testing (despite focusing mostly on anatomical imaging). However, when analyzing the description of datasets in Table 8.1, it emerged that information related to the number of patients/surgeries/healthcare centers involved in the dataset creation may be missing. This information could provide useful hints to be exploited by researcher when developing and testing DL algorithms (e.g. in terms of robustness to intra- and inter-patient variability) [103]. It is also worth noticing that the dataset numerosity (both in terms of images and patients) heavily varies from dataset to dataset (for each different clinical task). For example, in the MICCAI EndoVis dataset for small-bowel lesion localization, approximately 3600 images are given, while for early Barrett's cancer detection only 100 images are available.

Researchers are currently trying to overcome the DL shortcoming of requiring huge annotation datasets with unsupervised approaches where the problem of high dimensionality of the random variables to be modeled arise [33].

Even in presence of a sufficiently large labeled dataset, CNN training may not be trivial if the training labels are sparse, unbalanced or if there is not a consensus among health-operator annotations (e.g., in the definition of tumor margins). Specific weakly-supervised learning techniques, as multiple instance learning [104], may be used to address the problem of both temporal and spatial sparse labeling [105]. Solutions to face data unbalanced should be applied both at data and algorithm level [106], especially when training data are strongly unbalanced (i.e., number of positive cases $\ll$ number of negative cases). In order to improve the annotation

[2] https://grand-challenge.org/all_challenges/.

procedure, ranking algorithms [107, 108] can be used to sort the different responses of health-operator annotators, while evaluating the confidence level of the reported label.

In conclusion, to allow the actual integration of quantitative intra-operative image analysis into the actual clinical practice [109], the goal is developing adequate data-analysis technology to provide surgeons with quantitative support and effectively translate the technology into patient care workflow. SDS plays an important role in moving from (surgeon-specific) subjective to (computer-assisted) objective decision-making and from qualitative to quantitative assessment of surgical outcomes [2]. The integration of computer-aids will facilitate the surgeon's decision process and risk assessment, offering situation awareness, improved ergonomics and reduced cognitive workload.

References

1. Taylor, R.H., Menciassi, A., Fichtinger, G., Fiorini, P., Dario, P.: Medical robotics and computer-integrated surgery. In: Springer Handbook of Robotics, pp. 1657–1684. Springer (2016)
2. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nat. Biomed. Eng. **1**(9), 691 (2017)
3. Stewart, J.W., Akselrod, G.M., Smith, D.R., Mikkelsen, M.H.: Toward multispectral imaging with colloidal metasurface pixels. Adv. Mater. **29**(6), (2017)
4. Machida, H., Sano, Y., Hamamoto, Y., Muto, M., Kozu, T., Tajiri, H., Yoshida, S.: Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study. Endoscopy **36**(12), 1094–1098 (2004)
5. Emsley, J.W., Lindon, J.C.: NMR Spectroscopy Using Liquid Crystal Solvents. Elsevier (2018)
6. Abbou, C.C., Hoznek, A., Salomon, L., Olsson, L.E., Lobontiu, A., Saint, F., Cicco, A., Antiphon, P., Chopin, D.: Laparoscopic radical prostatectomy with a remote controlled robot. J. Urol. **197**(2), S210–S212 (2017)
7. Balmer, J.M., Yen, D.A.: The internet of total corporate communications, quaternary corporate communications and the corporate marketing internet revolution. J. Mark. Manag. **33**(1–2), 131–144 (2017)
8. Stoyanov, D.: Surgical vision. Ann. Biomed. Eng. **40**(2), 332–345 (2012)
9. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al.: Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Med. Image Anal. **17**(8), 974–996 (2013)
10. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
11. März, K., Hafezi, M., Weller, T., Saffari, A., Nolden, M., Fard, N., Majlesara, A., Zelzer, S., Maleshkova, M., Volovyk, M., et al.: Toward knowledge-based liver surgery: holistic information processing for surgical decision support. Int. J. Compu. Assist. Radiol. Surg. **10**(6), 749–759 (2015)
12. Moccia, S., Foti, S., Routray, A., Prudente, F., Perin, A., Sekula, R.F., Mattos, L.S., Balzer, J.R., Fellows-Mayle, W., De Momi, E., et al.: Toward improving safety in neurosurgery with an active handheld instrument. Ann. Biomed. Eng. **46**(10), 1450–1464 (2018)

13. Nosrati, M.S., Peyrat, J.M., Abinahed, J., Al-Alao, O., Al-Ansari, A., Abugharbieh, R., Hamarneh, G.: Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 324–331. Springer (2014)
14. Katić, D., Schuck, J., Wekerle, A.L., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy. Int. J. Comput. Assist. Radiol. Surg. **11**(6), 881–888 (2016)
15. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
16. Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al.: Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging **36**(6), 1231–1249 (2017)
17. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 1 (2018)
18. Moccia, S., De Momi, E., El Hadji, S., Mattos, L.S.: Blood vessel segmentation algorithm-sreview of methods, datasets and evaluation metrics. Comput. Methods Programs Biomed. **158**, 71–91 (2018)
19. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
20. Patel, V., Armstrong, D., Ganguli, M., Roopra, S., Kantipudi, N., Albashir, S., Kamath, M.V.: Deep learning in gastrointestinal endoscopy. Crit. Rev. Biomed. Eng. **44**(6) (2016)
21. Prasath, V.B.S.: Polyp detection and segmentation from video capsule endoscopy: a review. J. Imaging **3**(1) (2017)
22. de Souza, L.A., Palm, C., Mendel, R., Hook, C., Ebigbo, A., Probst, A., Messmann, H., Weber, S., Papa, J.P.: A survey on Barrett's esophagus analysis using machine learning. Comput. Biol. Med. (in press)
23. Zhang, J., Xia, Y., Xie, Y., Fulham, M., Feng, D.D.: Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. IEEE J. Biomed. Health Inform. **22**(5), 1521–1530 (2018)
24. Zhang, Y., Wirkert, S.J., Iszatt, J., Kenngott, H., Wagner, M., Mayer, B., Stock, C., Clancy, N.T., Elson, D.S., Maier-Hein, L.: Tissue classification for laparoscopic image understanding based on multispectral texture analysis. J. Med. Imaging **4**(1), 015,001–015,001 (2017)
25. Misawa, M., Kudo, S.E., Mori, Y., Takeda, K., Maeda, Y., Kataoka, S., Nakamura, H., Kudo, T., Wakamura, K., Hayashi, T., et al.: Accuracy of computer-aided diagnosis based on narrow-band imaging endocytoscopy for diagnosing colorectal lesions: comparison with experts. Int. J. Comput. Assist. Radiolo. Surg. 1–10 (2017)
26. Moccia, S., De Momi, E., Guarnaschelli, M., Savazzi, M., Laborai, A., Guastini, L., Peretti, G., Mattos, L.S.: Confident texture-based laryngeal tissue classification for early stage diagnosis support. J. Med. Imaging **4**(3), 034,502 (2017)
27. Freeman, W.T., Roth, M.: Orientation histograms for hand gesture recognition. In: International Workshop on Automatic Face and Gesture Recognition, vol. 12, pp. 296–301 (1995)
28. Magoulas, G.D.: Neuronal networks and textural descriptors for automated tissue classification in endoscopy. Oncol. Rep. **15**(4), 997–1000 (2006)
29. Kumar, S., Saxena, R., Singh, K.: Fractional fourier transform and fractional-order calculus-based image edge detection. Circuits Syst. Signal Process. **36**(4), 1493–1513 (2017)
30. Mukherjee, R., Manohar, D.D., Das, D.K., Achar, A., Mitra, A., Chakraborty, C.: Automated tissue classification framework for reproducible chronic wound assessment. BioMed Res. Int. **2014** (2014)
31. Karargyris, A., Bourbakis, N.: Wireless capsule endoscopy and endoscopic imaging: a survey on various methodologies presented. IEEE Eng. Med. Biol. Mag. **29**(1), 72–83 (2010)

32. Moccia, S., Wirkert, S.J., Kenngott, H., Vemuri, A.S., Apitz, M., Mayer, B., De Momi, E., Mattos, L.S., Maier-Hein, L.: Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Trans. Biomed. Eng. **158**(65), 2649–2659 (2018)
33. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
36. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561 (2015)
37. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications, pp. 179–187. Springer (2016)
38. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
39. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
40. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
42. Larsson, G., Maire, M., Shakhnarovich, G.: FractalNet: ultra-deep neural networks without residuals. arXiv:1605.07648 (2016)
43. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
44. Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A.W., et al.: Large-scale medical image annotation with crowd-powered algorithms. J. Med. Imaging **5**(3), 034,002 (2018)
45. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Trans. Med. Imaging **23**(7), 903–921 (2004)
46. Bi, L., Kim, J., Ahn, E., Feng, D.: Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv:1703.04197 (2017)
47. Sarker, M., Kamal, M., Rashwan, H.A., Banu, S.F., Saleh, A., Singh, V.K., Chowdhury, F.U., Abdulwahab, S., Romani, S., Radeva, P., et al.: SLSDeep: skin lesion segmentation based on dilated residual and pyramid pooling networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 21–29. Springer (2018)
48. Mirikharaji, Z., Hamarneh, G.: Star shape prior in fully convolutional networks for skin lesion segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention, pp. 737–745. Springer International Publishing, Cham (2018)
49. Ghosh, T., Li, L., Chakareski, J.: Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In: IEEE International Conference on Image Processing, pp. 3034–3038. IEEE (2018)
50. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: International Workshop on Machine Learning for Signal Processing, pp. 1–6. IEEE (2018)
51. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. **2017** (2017)

52. Brandao, P., Zisimopoulos, O., Mazomenos, E., Ciuti, G., Bernal, J., Visentini-Scarzanella, M., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., et al.: Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. J. Med. Robot. Res. **3**(02), 1840,002 (2018)
53. Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J.P., Eslami, A., Tombari, F., Navab, N.: Concurrent segmentation and localization for tracking of surgical instruments. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 664–672. Springer (2017)
54. Attia, M., Hossny, M., Nahavandi, S., Asadi, H.: Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 3373–3378. IEEE (2017)
55. García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S.: Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy, pp. 84–95. Springer (2016)
56. Milletari, F., Rieke, N., Baust, M., Esposito, M., Navab, N.: CFCM: segmentation via coarse to fine context memory. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention, pp. 667–674. Springer International Publishing, Cham (2018)
57. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J. Biomed. Health Inform. **21**(1), 65–75 (2017)
58. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122 (2015)
59. Bozkurt, A., Kose, K., Alessi-Fox, C., Gill, M., Dy, J., Brooks, D., Rajadhyaksha, M.: A multiresolution convolutional neural network with partial label training for annotating reflectance confocal microscopy images of skin. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention, pp. 292–299. Springer International Publishing, Cham (2018)
60. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv:1412.6806 (2014)
61. He, J.Y., Wu, X., Jiang, Y.G., Peng, Q., Jain, R.: Hookworm detection in wireless capsule endoscopy images with deep learning. IEEE Trans. Image Process. **27**(5), 2379–2392 (2018)
62. Mahmood, F., Durr, N.J.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Med. Image Anal. (2018)
63. Furukawa, R., Mizomori, M., Hiura, S., Oka, S., Tanaka, S., Kawasaki, H.: Wide-area shape reconstruction by 3D endoscopic system based on CNN decoding, shape registration and fusion. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 139–150. Springer (2018)
64. Oda, M., Roth, H.R., Kitasaka, T., Furukawa, K., Miyahara, R., Hirooka, Y., Goto, H., Navab, N., Mori, K.: Colon shape estimation method for colonoscope tracking using recurrent neural networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention, pp. 176–184. Springer International Publishing, Cham (2018)
65. Zhang, J., Xie, Y., Wu, Q., Xia, Y.: Skin lesion classification in dermoscopy images using synergic deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 12–20. Springer (2018)
66. Lopez, A.R., Giro-i Nieto, X., Burdick, J., Marques, O.: Skin lesion classification from dermoscopic images using deep learning techniques. In: International Conference on Biomedical Engineering, pp. 49–54. IEEE (2017)
67. Navarro, F., Conjeti, S., Tombari, F., Navab, N.: Webly supervised learning for skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 398–406. Springer (2018)

68. Pal, A., Chaturvedi, A., Garain, U., Chandra, A., Chatterjee, R.: Severity grading of psoriatic plaques using deep CNN based multi-task learning. In: International Conference on Pattern Recognition, pp. 1478–1483. IEEE (2016)
69. Mendel, R., Ebigbo, A., Probst, A., Messmann, H., Palm, C.: Barrett's esophagus analysis using convolutional neural networks. In: Bildverarbeitung für die Medizin 2017, pp. 80–85. Springer (2017)
70. Georgakopoulos, S.V., Iakovidis, D.K., Vasilakakis, M., Plagianakos, V.P., Koulaouzidis, A.: Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions. In: IEEE International Conference on Imaging Systems and Techniques, pp. 510–514. IEEE (2016)
71. Hong, J., Park, B.y., Park, H.: Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. In: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2892–2895. IEEE (2017)
72. Ribeiro, E., Uhl, A., Wimmer, G., Häfner, M.: Exploring deep learning and transfer learning for colonic polyp classification. Comput. Math. Methods Med. (2016)
73. Yuan, Y., Qin, W., Ibragimov, B., Han, B., Xing, L.: RIIS-DenseNet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 620–628. Springer (2018)
74. Aoki, T., Yamada, A., Aoyama, K., Saito, H., Tsuboi, A., Nakada, A., Niikura, R., Fujishiro, M., Oka, S., Ishihara, S., et al.: Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest. Endosc. (in press)
75. Fan, S., Xu, L., Fan, Y., Wei, K., Li, L.: Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. Phys. Med. Biol. **63**(16), 165,001 (2018)
76. Sekuboyina, A.K., Devarakonda, S.T., Seelamantula, C.S.: A convolutional neural network approach for abnormality detection in wireless capsule endoscopy. In: IEEE International Symposium on Biomedical Imaging, pp. 1057–1060. IEEE (2017)
77. Seguí, S., Drozdzal, M., Pascual, G., Radeva, P., Malagelada, C., Azpiroz, F., Vitrià, J.: Generic feature learning for wireless capsule endoscopy analysis. Comput. Biol. Med. **79**, 163–172 (2016)
78. Vasilakakis, M.D., Diamantis, D., Spyrou, E., Koulaouzidis, A., Iakovidis, D.K.: Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames. Evol. Syst. 1–13 (2018)
79. Itoh, T., Kawahira, H., Nakashima, H., Yata, N.: Deep learning analyzes helicobacter pylori infection by upper gastrointestinal endoscopy images. Endosc. Int. Open **6**(2), E139 (2018)
80. Yu, J.S., Chen, J., Xiang, Z., Zou, Y.X.: A hybrid convolutional neural networks with extreme learning machine for WCE image classification. In: IEEE International Conference on Robotics and Biomimetics, pp. 1822–1827. IEEE (2015)
81. Chen, H., Wu, X., Tao, G., Peng, Q.: Automatic content understanding with cascaded spatial-temporal deep framework for capsule endoscopy videos. Neurocomputing **229**, 77–87 (2017)
82. Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., Maier, A.: Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. Sci. Rep. **7**(1), 11,979 (2017)
83. Xu, T., Zhang, H., Huang, X., Zhang, S., Metaxas, D.N.: Multimodal deep learning for cervical dysplasia diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 115–123. Springer (2016)
84. Zou, Y., Li, L., Wang, Y., Yu, J., Li, Y., Deng, W.: Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network. In: IEEE International Conference on Digital Signal Processing, pp. 1274–1278. IEEE (2015)
85. Zhou, T., Han, G., Li, B.N., Lin, Z., Ciaccio, E.J., Green, P.H., Qin, J.: Quantitative analysis of patients with celiac disease by video capsule endoscopy: a deep learning method. Comput. Biol. Med. **85**, 1–6 (2017)

86. Du, X., Kurmann, T., Chang, P.L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J.D., Stoyanov, D.: Articulated multi-instrument 2D pose estimation using fully convolutional networks. IEEE Trans. Med. Imaging (2018)
87. Colleoni, E., Moccia, S., Du, X., De Momi, E., Stoyanov, D.: Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robot. Autom. Lett. **4**(3), 2714–2721 (2019)
88. Aubreville, M., Stoeve, M., Oetter, N., Goncalves, M., Knipfer, C., Neumann, H., Bohr, C., Stelzle, F., Maier, A.: Deep learning-based detection of motion artifacts in probe-based confocal laser endomicroscopy images. Int. J. Comput. Assist. Radiol. Surg. (in press)
89. Sano, Y., Emura, F., Ikematsu, H.: Narrow-band imaging. In: Colonoscopy: Principles and Practice, 2nd edn., pp. 514–526 (2009)
90. Li, Q., He, X., Wang, Y., Liu, H., Xu, D., Guo, F.: Review of spectral imaging technology in biomedical engineering: achievements and challenges. J. Biomed. Opt. **18**(10), 100,901–100,901 (2013)
91. Zeng, C., King, D.J., Richardson, M., Shan, B.: Fusion of multispectral imagery and spectrometer data in UAV remote sensing. Remote Sens. **9**(7), 696 (2017)
92. Wirkert, S.J., Vemuri, A.S., Kenngott, H.G., Moccia, S., Götz, M., Mayer, B.F., Maier-Hein, K.H., Elson, D.S., Maier-Hein, L.: Physiological parameter estimation from multispectral images unleashed. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 134–141. Springer (2017)
93. Moccia, S., Vanone, G.O., De Momi, E., Laborai, A., Guastini, L., Peretti, G., Mattos, L.S.: Learning-based classification of informative laryngoscopic frames. Comput. Methods Programs Biomed. **158**, 21–30 (2018)
94. McLaren, B., Ashley, K.: Helping a CBR program know what it knows. In: Case-Based Reasoning Research and Development, pp. 377–391 (2001)
95. Obermeyer, Z., Emanuel, E.J.: Predicting the futurebig data, machine learning, and clinical medicine. N. Engl. J. Med. **375**(13), 1216 (2016)
96. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: a tutorial overview. Neuroimage **45**(1), S199–S209 (2009)
97. Zhang, Q., Wu, Y.N., Zhu, S.: Interpretable convolutional neural networks. CoRR arXiv:1710.00935 (2017)
98. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Trans. Image Process. **27**(5), 2368–2378 (2018)
99. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(05), 557–570 (2002)
100. Cocos, A., Qian, T., Callison-Burch, C., Masino, A.J.: Crowd control: effectively utilizing unscreened crowd workers for biomedical data annotation. J. Biomed. Inform. **69**, 86–92 (2017)
101. Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., Stock, C., Heim, E., Götz, M., Wirkert, S., Kenngott, H., et al.: Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 616–623. Springer (2016)
102. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int. J. Comput. Assist. Radiol. Surg. 1–9 (2018)
103. Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P.M., Bogunovic, H., Landman, B.A., Maier, O., Menze, B., et al.: How to exploit weaknesses in biomedical challenge design and organization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 388–395. Springer (2018)
104. Moccia, S., Mattos, L.S., Patrini, I., Ruperti, M., Poté, N., Dondero, F., Cauchy, F., Sepulveda, A., Soubrane, O., De Momi, E., et al.: Computer-assisted liver graft steatosis assessment via learning-based texture analysis. Int. J. Comput. Assist. Radiol. Surg. 1–11 (2018)
105. Bernardini, M., Romeo, L., Misericordia, P., Frontoni, E.: Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. IEEE J. Biomed. Health Inform. (2019)

106. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets **2**, 42–47 (2012)
107. Heikkilä, T., Dalgaard, L., Koskinen, J.: Designing autonomous robot systems-evaluation of the r3-cop decision support system approach. In: SAFECOMP 2013-Workshop DECS (ERCIM/EWICS Workshop on Dependable Embedded and Cyber-physical Systems) of the 32nd International Conference on Computer Safety, Reliability and Security, p. NA (2013)
108. Hansen, P., Ombler, F.: A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. J. Multi-Criteria Decis. Anal. **15**(3–4), 87–107 (2008)
109. D'Haese, P.F., Konrad, P.E., Pallavaram, S., Li, R., Prassad, P., Rodriguez, W., Dawant, B.M.: CranialCloud: a cloud-based architecture to support trans-institutional collaborative efforts in neurodegenerative disorders. Int. J. Comput. Assist. Radiol. Surg. **10**(6), 815–823 (2015)

Chapter 9
Convolutional Neural Networks for 3D Protein Classification

Loris Nanni, Federica Pasquali, Sheryl Brahnam, Alessandra Lumini and Apostolos Axenopoulos

Abstract The main goal of this chapter is to develop a system for automatic protein classification. Proteins are classified using CNNs trained on ImageNet, which are tuned using a set of multiview 2D images of 3D protein structures generated by Jmol, which is a 3D molecular graphics program. Jmol generates different types of protein visualizations that emphasize specific properties of a protein's structure, such as a visualization that displays the backbone structure of the protein as a trace of the C_α atom. Different multiview protein visualizations are generated by uniformly rotating the protein structure around its central X, Y, and Z viewing axes to produce 125 images for each protein. This set of images is then used to fine-tune the pretrained CNNs. The proposed system is tested on two datasets with excellent results. The MATLAB code used in this chapter is available at https://github.com/LorisNanni.

9.1 Introduction

Comparing protein structures is one of the most important tasks in structural biology: comparisons are essential, for instance, in inferring protein evolution and in understanding the relationship between protein structure and function. Comparing

L. Nanni (✉) · F. Pasquali
Department of Information Engineering, University of Padua, Via Gradenigo 6, 35131 Padova, Italy
e-mail: loris.nanni@unipd.it

S. Brahnam
Department of Information Technology and Cybersecurity, Missouri State University, Glass Hall, Room 387, 901 S. National, Springfield, MO 65804, USA

A. Lumini
DISI - Department of Computer Science and Engineering, Università di Bologna, Via Sacchi 3, 47521 Cesena, Italy
e-mail: alessandra.lumini@unibo.it

A. Axenopoulos
University of Thessaly, Volos, Greece

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_9

pairs of protein structures often begins with an alignment process accomplished by considering the equivalence between pairs of amino acid residues. Alignment is then followed by a search for the type of geometrical transformation that minimizes the distance between the residues [1]. Finally, a dissimilarity measure is computed that ideally produces a single number that distinguishes between related and nonrelated structure pairs. An example of such a measure is the Root Mean Square Deviation (RMSD) calculated between the superimposed C_α chains.

Many protein structural alignment methods have been proposed. Some popular methods include DALI (Distance matrix ALIgnment) [2], CE (Combinatorial Extension) [3] and FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists) [4], all of which are available on the protein databank (PDB) website [5]. These alignment methods, however, often have difficulties when comparing protein structures that are markedly dissimilar. This difficulty is based on determining a single optimal alignment of functional similarity and/or tertiary structure similarity [6] and is magnified whenever the sequence identity of the protein is less than or equal to 20%, the point at which structural differences become very large [7]. Another problem has to do with using RMSD as a standalone dissimilarity measure. Though RMSD depends on the length of the alignment, a greater number of aligned positions do not necessarily result in a smaller RMSD value [4, 8]. To compensate for this lack of refinement, many scoring functions have been proposed to complement RMSD, especially for classification and retrieval tasks. Two examples of complementary dissimilarity measures include the application of Z-score statistics [9] and of template modeling (TM) scores [8], which apply a weight based on the length of the aligned structure.

Unlike these local similarity measures, global representations apply feature extraction to 3D geometrical structures [10–16]. Similarity between structures in global approaches is not based on alignment, as with local approaches, but rather on a comparison of feature vectors that represent the structures. In [13], for example, 3D Zernike descriptors are used to represent the 3D geometrical surface of the protein structures, in [55] 2D Polar-Fourier coefficients and 2D Krawtchouk moments are applied resulting in a rotation-invariant descriptor vector, and in [16] wavelet-based protein descriptors are proposed. In [10] and [15], protein structures are represented as open curves from which descriptors based on Gauss integral vectors are extracted, and in [11] protein structures are represented by a symmetric interaction matrix that contains parameters of the relationship between secondary structure elements. In [12] and [14], a tableau containing the encoded orientation of the secondary structural elements describes protein structure. In most global shape matching approaches, the 3D molecules are treated as a rigid object. To address the flexibility of protein structures, some approaches [53, 54] have been proposed that transform the Euclidean metrics into a metric space where the pairwise distances between points of the 3D object surface are invariant to deformations of the 3D object (e.g., geodesic distances, inner distances, or diffusion distances). Though far simpler, the vector representations extracted from global approaches are nonetheless capable of representing large and complicated protein structures.

It should be noted that both local alignment-based methods and global feature-based approaches explicitly exploit geometrical information. The problems encountered in local conventional methods are all geometric in nature. One recently proposed method for overcoming the difficulties already described in the alignment process and the instability of measurements is the view-based approach to protein structure comparison introduced in [17, 18]. The view-based approach is based on a set of 2D multiviews of 3D molecular visualization images. The basic idea behind this method derives from that fact that in many studies of protein analysis, manual inspection of protein structures relies on visualization software [19], where 3D structures are inspected after they have been projected from multiple viewpoints onto the 2D plane of the computer screen using several types of representation: Ball&Stick, Backbone, Ribbons, Cartoons, etc. Researchers recognize that each of these protein representations provides access to different types of information about a protein's structure. For example, the overall structure of the protein can be extracted from a backbone representation, and the direction of the protein structural elements is evident in cartoon visualizations. Rocket visualization highlights the composition of the secondary structures, and directed cylinders and arrows visualize the helical and stranded structures of proteins. The developers of the view-based approach recognized that the information contained in these visualizations could be used in machine-based comparisons of proteins. So rather than directly using 3D geometrical information, the view-based approach uses a set of multiview images generated by rotating the protein structures in a 3D molecular graphics program. An advantage offered by 3D molecular graphics software is its ability to synthesize different types of images that emphasize different protein properties.

After collecting a set of multiview images, robust state-of-the-art image descriptors can then be extracted from each image. In [18] the authors obtained a feature vector based on Local Binary Patterns (LBP) [20] and other handcrafted texture descriptors, after which a protein subspace was generated by applying Principal Component Analysis (PCA) to the set of feature vectors. The final step in the view-based approach was to characterize the similarity between protein structures using the canonical angles θi between the corresponding subspaces, a measure of similarity known as the Mutual Subspace Method (MSM) [21]. MSM-based methods [22–24] are known for their robust performance in classifying complicated yet similar 3D shapes (e.g., in facial [25, 26] and hand shape recognition [27]). However, there is a problem using methods based on MSM. Although it captures the overall structural similarity between two proteins, it is inadequate for distinguishing between two different protein types that have a similar shape. Because MSM is the computation of the distance between points on a Grassmann manifold, an improvement in discriminative power can be obtained by extracting more information from each corresponding subspace using statistical analysis on the Grassmann manifold.

Using texture features such as LBP is not the only way to represent images. In the last decade, deep learning techniques have revolutionized the field of pattern classification. Convolutional Neural Networks (CNNs), for example, have significantly advanced image recognition, including biomedical image classification. Deep learners, such as CNN, however, usually require large datasets and a great deal of

computational power to achieve good results. One way of circumventing a need for large datasets and such intensive training is to fine-tune or tune a pretrained deep learner (i.e., one that has already been trained on a large dataset of images for some other image classification problem) by subjecting it to additional training so that it works well with an image dataset (even one that is small) representing an entirely different task.

In this chapter, we report our first experiments classifying proteins using CNNs pretrained on ImageNet [28] and then fine-tuned using a set of multiview 2D images of 3D protein structures generated using Jmol. We test two CNN architectures: AlexNet [29] and GoogleNet [30].

The main objective of this chapter is to experimentally develop an automatic protein classification system by extracting different types of descriptors from the protein visualizations. Ensembles composed of these descriptors are then tested on two benchmark datasets [18]: (1) a seven-class classification according to the protein classification scheme used in the SCOP database, and (2) a dataset where the proteins that belong to the membrane class of the SCOP database are divided into five folds.

9.2 Methods

In this work, we test the performance of AlexNet [29] and GoogleNet [30] trained on ImageNet [28] for protein classification. The CNNs are fine-tuned with images of protein structure generated by Jmol. The CNNs are described in Sect. 9.2.2, and the generated Jmol images in Sect. 9.2.1.

Standard techniques are also adopted for comparing CNN performance and for building an ensemble of classifiers (that are also combined with the CNNs). The feature sets used in the standard techniques are briefly described in Sects. 9.2.3–9.2.7. The standard techniques train an ensemble of Support Vector Machine (SVM) classifiers, where a single SVM is trained on each feature set.

9.2.1 Generation of Multiview Protein Images

Today there are many sophisticated 3D molecular graphics programs (see [31] for a comprehensive review) that allow users to visualize the structure of proteins. In this work, we used Jmol [32] to generate different types of 3D visualizations that emphasize specific properties of protein structure: backbone visualization, which displays the backbone structure of the protein as a trace of the C_α atom; ribbon visualization, which displays the backbone structure as a smooth ribbon shape showing the helical shape of the α-helix structure; rocket visualization, which contains information on the secondary structure of the protein by displaying the secondary structure of α-helices as directed cylinders and β-sheets as arrows, while random coils are visualized as strings; and cartoon visualization, which displays α-helices as ribbons, β-sheets as

directed ribbons, and random coils as strings. Because each visualization highlights different protein characteristics, these visualizations can be used to produce more elaborate protein descriptors. As already mentioned, these multiview visualizations are synthesized by uniformly rotating the protein structure around its central X, Y, and Z viewing axes. Generation of uniform rotation angles can be regarded as equivalent to producing a uniform distribution of points on the surface of a sphere, where each point is the viewpoint for protein visualization.

9.2.2 Convolutional Neural Networks

CNNs is a deep learning architecture that has been extensively studied [33]. CNNs produce accurate and generalizable models and have achieved state-of-the-art performance in many pattern recognition problems.

A CNN is a multi-layered image classification network that incorporates spatial context and weight sharing between pixels to learn the optimal image features for a classification task. Different types of layers are used to build a CNN: convolutional, pooling, and fully-connected layers, for instance, whose weights are trained with the backpropagation algorithm using a large labeled dataset. In cases where the training set is insufficiently large enough to preform training from scratch, transfer learning [34] has proven useful. Because CNNs have considerable generalization power [35], a pretrained model can be used in two ways: (1) as a feature extractor to obtain descriptors learned using an image dataset and (2) as a classifier for a different image classification problem by tuning the weights of the network to the new problem.

Many CNN architectures have been proposed: LeNet [36], AlexNet [29], VGGNet [37], GoogleNet [30] and ResNet [38] being some of the most famous, with the CNNs in this list ranging from one of the simplest and lightest CNNs (LeNet) to one of the most complex and deepest architectures (ResNet). For the protein classification task, we tested many different networks but obtained the best performance using AlexNet [29]. This CNN was proposed in 2012 and won the ImageNet ILSVRC challenge that year. AlexNet is composed of both stacked and connected layers: five convolutional layers followed by three fully-connected layers, with some max-pool layers in the middle and a rectified linear unit nonlinearity for each convolutional and fully connected layer.

We obtained the best performance with AlexNet using the following tuning parameters: the maximum number of epochs for training set to 20, the mini-batch size set to {30, 70} and a fixed learning rate of 0.001 and 0.0001. For building some additional poses to be trained with AlexNet, we used random reflection, where each image was reflected horizontally with 50% probability.

9.2.3 Descriptor for Primary Representation: Quasi Residue Couple (QRC)

Inspired by Chou's quasi-sequence-order model and Yuan's Markov chain model [39], QRC [40] is a method for extracting features from the primary sequence of a protein [41]. The original residue couple model was designed to represent the information contained in both the amino acid composition (AAC) and the order of the amino acids in the protein sequences. The QRC descriptor is obtained by selecting a physicochemical property d and combining its values with each nonzero entry in the residue couple. A parameter m represents the order of the residue couple model, with values of $m \leq 3$ deemed sufficient for representing a sequence.

The QRC model for a physicochemical property d, can be represented as:

$$\mathrm{QRC}_m^d(k) = \frac{1}{N-m} \sum_{n=1}^{N-m} H_{i,j}(n, n+m, d) + H_{j,i}(n+m, n, d),$$

where $i, j \in [1, \ldots, 20]$ are the twenty different amino acids; $k = j + 20(i-1)$, N is the length of the protein, the function $index(i, d)$ returns the value of the property d for the amino acid i, and the function $H_{i,j}(a, b, d) = index(i, d)$, if $p_a = i$ and $p_b = j$, otherwise $H_{i,j}(a, b, d) = 0$.

In the experimental section, QRC^d features are extracted for m in the range of 1–3 and concatenated into a 1200-dimensional vector. A total of twenty-five physicochemical properties are randomly selected to create an ensemble of *QRC* descriptors.

9.2.4 Descriptor for Primary Representation: Autocovariance Approach (AC)

AC [42], based on autocovariance as the name suggests, is a sequence-based variant of Chou's pseudo amino acid composition (PseAAC) [43]. AC extracts a set of PseAAC-based features from a given protein, which is the concatenation of the twenty standard AAC values, along with m values reflecting the effect of the sequence order. The parameter m (set to 20 in this work) indicates the maximum distance between two amino acids i and j.

Given a protein $P = (p_1, p_2, \ldots, p_N)$, and fixing physicochemical property d, the AC descriptor ($\mathrm{AC}^d \in \Re^{20+m}$) can be defined as:

$$\mathrm{AC}^d(i) = \begin{cases} h(i)/N & i \in [1, \ldots, 20] \\ \sum_{k=1}^{N-i+20} \frac{(index(\mathrm{p}_k, d) - \mu_d) \cdot (index(\mathrm{p}_{k+i-20}, d) - \mu_d)}{\sigma_d \cdot (N-i+20)} & i \in [21, \ldots, 20+m] \end{cases} \tag{9.1}$$

where the function $index(i, d)$ returns the value of the property d for the amino acid i, and the function $h(i)$ counts the number of occurrences of a given amino acid in a protein sequence. The normalization factors μ_d and σ_d are the mean and the variance of d on the twenty amino acids:

$$\mu_d = \frac{1}{20}\sum_{i=1}^{20} index(i, d), \sigma_d = \frac{1}{20}\sum_{i=1}^{20} (index(i, d) - \mu_d)^2. \tag{9.2}$$

Here twenty-five random physicochemical properties are selected to create an ensemble of twenty-five AC descriptors.

9.2.5 *Matrix Representation for Proteins: Position Specific Scoring Matrix (PSSM)*

PSSM [44] is a matrix representation extracted from a group of sequences previously aligned by structural or sequence similarity. PSSM is calculated using PSI-BLAST, an application that compares PSSM profiles for detecting remotely related homologous proteins or DNA.

The following parameters are considered:

1. Position: the index of each amino acid residue in a sequence after multiple sequence alignments;
2. Probe: a group of typical sequences of functionally related proteins that have already been aligned by structural similarity or sequence;
3. Profile: a matrix of 20 columns that correspond to the 20 amino acids;
4. Consensus: the sequence of amino acid residues that are most similar to all the alignment residues of probes at each position (the consensus sequence is calculated by selecting the highest score at each position in the profile).

The PSSM representation for a given protein of length N is a matrix with dimension $N \times 20$:

$$\text{PSSM} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,20} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ S_{N,1} & S_{N,2} & \cdots & S_{N,20} \end{bmatrix}, \tag{9.3}$$

where $S_{i,j}$ represents the occurrence probability of amino acid j at position i of the protein sequence. The rows in the matrix represent the positions of the sequence, and the columns represent the 20 original amino acids.

The elements of PSSM(i, j) are calculated as:

$$\text{PSSM}(i,\ j) = \sum_{k=1}^{20} w(i,k) \times Y(j,k), i = (1,,N);\ j = (1,,20), \tag{9.4}$$

where $w(i,k)$ is the ratio between the frequency of the kth amino acid at the position i of the probe and the total number of probes, and $Y(j,k)$ is the value of Dayhoff's mutation matrix between the jth and kth amino acids. It should be noted that $Y(j,k)$ is a substitution matrix (a matrix that describes the rate at which one character in the protein changes into another over time).

PSSM scores are typically either positive or negative integers. Small values of PSSM$(i,\ j)$ indicate weakly conserved positions (meaning that the given amino acid occurs more frequently in the alignment than expected by chance), and large values indicate strongly conserved positions (meaning that the given amino acid occurs less frequently than expected). An element of a PSSM profile can be used to approximate the occurrence probability of the corresponding amino acid at a specific position.

9.2.6 Matrix Representation for Proteins: 3D Tertiary Structure (DM)

The DM representation is based on the distances between atoms and between residues in a PDB structure. DM creates a heat map displaying inter-residue distances. If the size of the map exceeds 250 it is resized to 250×250 to reduce the computation time for extracting the features. As is the case with the other protein matrix representations described in this paper, DM is treated as a grayscale image that is used to extract texture descriptors.

9.2.7 Matrix-Based Descriptors: Texture Descriptors

A protein matrix representation can be treated as an image from which robust texture descriptors can be extracted. In this work, we combine texture descriptors (see Table 9.1) by training each one with a separate SVM. Sets of SVMs are then combined by sum rule.

9.3 Experiments

The following two datasets were used to experimental build and test our system:

1. Dataset 700: this is a dataset where proteins are classified into seven classes according to the following protein classification scheme of the SCOP database [51]: (1) α-proteins (containing mainly α-helices); (2) β-proteins (containing

Table 9.1 Summary of texture descriptors and parameter settings

Label	Parameters	Source
LBP	Uniform LBP with two settings configurations (radius, number of neighbors P): (1, 8) and (2, 16)	[45]
WLD	Weber Law Descriptor code computed within a 3×3 block with the following parameter configurations: BETA = 5, ALPHA = 3, and number of neighbors = 8	[46]
CLBP	Completed LBP with two configurations (R,P): (1, 8) and (2, 16)	[47]
RIC	Multiscale rotation invariant co-occurrence of adjacent LBP with $R \in \{1, 2, 4\}$	[48]
MORPH	A set of MORHphological features, which is a set of measures that includes such features as the aspect ratio, number of objects, area, perimeter, eccentricity, and other measures extracted from a segmented version of the image	[49]
HASH	Default values of the heterogeneous auto-similarities of characteristics features. HASH is an LBP variant that models linear and non-linear feature dependencies	[50]

mainly β-sheets); (3) α/β-proteins (containing both α and β structures where the β-sheets are parallel); (4) α + β-proteins (containing both α and β structures where the β-sheets are anti-parallel); (5) multi-domain proteins that have multi-functions; (6) membrane and cell surface proteins; and (7) small proteins. This resulted in a total of 700 proteins, where 100 were randomly selected from each class (having at most 20% sequence identity) from the Astral dataset [52]. A 10-fold cross-validation protocol was used where 10 proteins from each class formed the testing set while the remainder formed the training set. This protocol was repeated 10 times for each experiment.

2. Dataset 95: this dataset was built by collecting protein structures from the Astral SCOP database that had no more than 10% sequence identity between the structures in the membrane class. The folds were sorted by the size of their populations so that each fold contained a minimum of 5 and a maximum of 50 proteins. Folds originally containing more than 50 proteins were reduced by randomly selecting up to 50 proteins. This procedure produced 95 proteins involving five folds: (1) f.1 toxins' membrane translocation domains, (2) f.17 transmembrane helix hairpin, (3) f.21 heme-binding four-helical bundle, (4) f.23 single transmembrane helix, and (5) f.4 transmembrane beta-barrels. Because the number of proteins in each fold was not balanced, stratified random sampling was used to feed the data into 10-fold cross-validations.

The first set of experiments were designed to evaluate the performance of the features proposed in this work using accuracy as the performance indicator. In Tables 9.2 and 9.3, we report the performance obtained using different values of batch size (BS) and learning rates (LR) for the CNNs. If a cell contains the value 'FUS_X', then what is meant is that the classifiers were combined by average rule and the CNN named 'X' was trained with BS = {30, 70} and LR = {0.001, 0.0001}.

Table 9.2 Performance in the 95-dataset

AlexNet—Dataset 95	BS = 30	BS = 70
LR = 0.001	0.800	0.822
LR = 0.0001	0.822	0.833
FUS_Alex 0.833		
GoogleNet – Dataset 95	BS = 30	BS = 70
LR = 0.001	0.788	0.833
LR = 0.0001	0.800	0.800
FUS_Googlenet 0.833		

Table 9.3 Performance in the 700-dataset

AlexNet—Dataset 700	BS = 30	BS = 70
LR = 0.001	0.565	0.564
LR = 0.0001	0.581	0.575
FUS_Alex 0.584		
GoogleNet—Dataset 700	BS = 30	BS = 70
LR = 0.001	0.584	0.548
LR = 0.0001	0.575	0.567
FUS_GoogleNet 0.567		

When we combined (with average sum rule) FUS_AlexNet and FUS_GoogleNet, the performance did improve (Dataset 95: 0.800; Dataset 700: 0.574).

In Tables 9.4 and 9.5, we report the performance obtained using standard protein descriptors and our proposed ensemble. The acronym TXT represents an ensemble of SVMs trained with the texture descriptors extracted from PSSM and DM combined by average rule. FUS_Alex + TXT is the fusion by sum rule between TXT and FUS_Alex (before fusion the scores of both approaches are normalized to mean 0 and standard deviation 1). In Table 9.5 we report results using another performance indicator: the area under the Receiver Operating Characteristic curve (AUC).

Table 9.4 Ensemble proposed—accuracy

Accuracy	TXT	AC	QRC	FUS_Alex	FUS_Alex + TXT
700-dataset	0.628	0.471	0.432	0.584	0.613
95-dataset	0.822	0.711	0.677	0.833	0.833

Table 9.5 Ensemble proposed—area under the ROC curve

AUC	TXT	AC	QRC	FUS_Alex	FUS_Alex + TXT
700-dataset	88.39	80.83	78.05	88.09	90.06
95-dataset	93.60	84.00	82.37	93.74	95.42

Table 9.6 Comparison with literature

Dataset 700	
[18] only backbone image extracted from protein	0.603
[18] best approach	0.694
[3]	0.491
[4]	0.531
[8]	0.640
FUS_Alex + TXT	0.613
Dataset 95	
[18] best approach	0.884
[8]	0.863
FUS_Alex + TXT	0.833

Clearly, AC and QRC, both based on amino-acid sequence, produce the worst results. The best performance (considering both performance indicators) is obtained by FUS_Alex + TXT.

Finally, in Table 9.6, we compare our approach with the literature (using accuracy as the performance indicator).

Our approach obtains a performance that competes well with the best in the literature, following an approach similar to [18] that uses only the backbone image for 3D protein representation. Even though our results are not superior to those reported in the literature, we have nonetheless clearly achieved our aim on the two datasets, viz., to show that a pretrained CNN can be tuned with 2D projection of 3D proteins.

9.4 Conclusion

In this chapter, our aim has been to show that the backbone protein image projected on a 2D plane can be used to tune a CNN. Experiments on two datasets, where we obtain a performance that is comparable with literature, shows that we have achieved our aim.

Using Jmol to generate sets of visualization images base on Backbone visualization (which displays the backbone structure of the protein as a trace of the C_α atom). Multiview protein visualization images are synthesized by uniformly rotating the protein structure around its central X, Y, and Z viewing axes to generate 125 images for each protein.

Several future works are planned:

- To test other available pretrained CNN on additional protein datasets;
- To combine CNN features with handcrafted features extracted from 2D images;
- To tests different 3D protein representations.

Acknowledgements We would like to acknowledge the support that NVIDIA provided us through the GPU Grant Program. We used a donated TitanX GPU to train CNNs used in this work.

References

1. Marti-Renom, M., Capriotti, E., Shindyalov, I., Bourne, P.: Structure comparison and alignment. In: Gu, J., Bourne, P.E. (eds.) Structural Bioinformatics, pp. 397–418. Wiley-Blackwell, Hoboken, NJ (2009)
2. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**, 123–138 (1993)
3. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11**(9), 739–747 (1998)
4. Ye, Y., Godzik, A.: Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics **19**(Suppl 2), ii246-ii255 (2003)
5. Berman, H.M., et al.: The protein data bank. Nucleic Acids Res. **28**(1), 235–242 (2000)
6. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22), 4673–4680 (1994)
7. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. The EMBO J. **5**(4), 823–826 (1986)
8. Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. **33**(7), 2302–2309 (2005)
9. Prlic, A., et al.: Precalculated protein structure alignments at the RCSB PDB website. Bioinformatics **26**, 2983–2985 (2010)
10. Røgen, P.: Evaluating protein structure descriptors and tuning Gauss integral based descriptors. J. Phys. Condens. Matter **17**, 1523–1538 (2005)
11. Zhou, X., Chou, J., Wong, S.T.C.: Protein structure similarity from principle component correlation analysis. BMC Bioinform. **7**(40) (2006)
12. Konagurthu, A.S., Stuckey, P.J., Lesk, A.M.: Structural search and retrieval using a tableau representation of protein folding patterns. Bioinformatics **24**(5), 645–651 (2008)
13. Sael, L., et al.: Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins **72**, 1259–1273 (2008)
14. Stivala, A., Wirth, A., Stuckey, P.J.: Tableau-based protein substructure search using quadratic programming. BMC Bioinform. **10**(1), 153 (2009)
15. Harder, T., Borg, M., Boomsma, W., Røgen, P., Hamelryck, T.: Fast large-scale clustering of protein structures using Gauss integrals. Bioinformatics, 510–515 (2012)
16. Mirceva, G., Cingovska, I., Dimov, Z., Davcev, D.: Efficient approaches for retrieving protein tertiary structures. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(4), 1166–1179 (2012)
17. Suryanto, C.H., Jiang, S., Fukui, K.: Protein structure similarity based on multi-view images generated from 3D molecular visualization. In: Presented at the 21st International Conference on Pattern Recognition (2012)
18. Suryanto, C.H., Saigo, H., Fukui, K.: Structural class classification of 3d protein structure based on multi-view 2d images. IEEE/ACM Trans. Comput. Biol. Bioinform. **15**(1), 286–299 (2015)
19. Bottomley, S., Helmerhorst, E.: Molecular visualization. In: Structural Bioinformatics, 2nd edn., pp. 237–268. Wiley-Blackwell, Hoboken, NJ (2009)
20. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distribution. Pattern Recogn. Lett. **29**(1), 51–59 (1996)
21. Maeda, K.: From the subspace methods to the mutual subspace method. In: Computer Vision. Studies in Computational Intelligence, vol. 285, pp. 135–156. Springer, Berlin and Heidelberg (2010)

22. Fukui, K., Stenger, B., Yamaguchi, O.: A framework for 3d object recognition using the kernel constrained mutual subspace method. In: Computer Vision—ACCV 2006. Lecture Notes in Computer Science, no. 3852, pp. 315–332. Springer, Berlin and Heidelberg (2006)
23. Fukui, K., Yamaguchi, O.: The kernel orthogonal mutual subspace method and its application to 3d object recognition. In: Computer Vision—ACCV 2007. Lecture Notes in Computer Science, vol. 4844, pp. 467–476. Springer, Berlin and Heidelberg (2007)
24. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1005–1018 (2007)
25. Fukui, K., Yamaguchi, O.: Face recognition using multiviewpoint patterns for robot vision. In: Presented at the 11th International Symposium of Robotics Research (2003)
26. Fukui, K., Maki, A.: Difference subspace and its generalization for subspace-based methods. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2164–2177 (2015)
27. Ohkawa, Y., Fukui, K.: Hand-shape recognition using the distributions of multi-viewpoint image sets. IEICE Trans. Inf. Syst. **E95-D**(6), 1619–1627 (2012)
28. Russakovsky, O., Deng, J., Su, H.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 1097–1105. Curran Associates Inc., Red Hook, NY (2012)
30. Szegedy, C., et al.: Going deeper with convolutions. In: Presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
31. O'Donoghue, S.I., et al.: Visualization of macromolecular structures. Nat. Methods **7**(3 Suppl), S42–S55 (2010)
32. Hanson, R.M.: Jmol—A paradigm shift in crystallographic vi-sualization. J. Appl. Crystallogr. **45**(5), 1250–1260 (2010)
33. Gua, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018)
34. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Cornell University (2014). arXiv:1411.1792
35. Nanni, L., Ghidoni, S., Brahnam, S.: Handcrafted versus non-handcrafted features for computer vision classification. Pattern Recogn. **71**, 158–172 (2017)
36. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceeding IEEE **86**(11), 2278–2323 (1998)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Cornell University (2014). arXiv:1409.1556v6
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV (2016)
39. Guo, J., Lin, Y., Sun, Z.: A novel method for protein subcellular localization: Combining residue-couple model and SVM. In: Presented at the Proceedings of 3rd Asia-Pacific Bioinformatics Conference, Singapore (2005)
40. Nanni, L., Lumini, A.: An ensemble of K-Local Hyperplane for predicting Protein-Protein interactions. Bioinformatics **22**(10), 1207–1210 (2006)
41. Nanni, L., Brahnam, S., Lumini, A.: High performance set of PseAAC descriptors extracted from the amino acid sequence for protein classification. J. Theor. Biol. **266**(1), 1–10 (2010)
42. Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., Li, M.L.: Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. **259**(2), 366–372 (2009)
43. Chou, K.-C.: Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics **6**, 262–274 (2009)
44. Gribskov, M., McLachlan, A.D., Eisenberg, D.: Profile analysis: detection of distantly related proteins. In: Presented at the Proceedings of the National Academy of Sciences (PNAS) (1987)

45. Ojala, T., Pietikainen, M., Maeenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
46. Chen, J., et al.: WLD: A robust local image descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1705–1720 (2010)
47. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. **19**(6), 1657–1663 (2010)
48. Nosaka, R., Fukui, K.: HEp-2 cell classification using rotation invariant co-occurrence among local binary patterns. Pattern Recogn. Bioinform. **47**(7), 2428–2436 (2014)
49. Strandmark, P., Ulén, J., Kahl, F.: HEp-2 staining pattern classification. In: Presented at the International Conference on Pattern Recognition (ICPR2012) (2012). https://lup.lub.lu.se/search/ws/files/5709945/3437301.pdf
50. San Biagio, M., Crocco, M., Cristani, M., Martelli, S., Murino, V.: Heterogeneous auto-similarities of characteristics (hasc): exploiting relational information for classification. In: Presented at the IEEE Computer Vision (ICCV13), Sydneys, Australia (2013)
51. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**(4), 536–540 (1995)
52. Fox, N.K., Brenner, S.E., Chandonia, J.-M.: SCOPe: Structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acid Res. **42**(Database), D304–09 (2014)

Part IV
Ethical Considerations

Chapter 10
From Artificial Intelligence to Deep Learning in Bio-medical Applications

Olga Lucia Quintero Montoya and Juan Guillermo Paniagua

Abstract Since their introduction in late 80s, convolutional neural networks and auto-encoder architectures have shown to be powerful for automatic feature extraction and information simplification. Using convolution kernels from image processing in 2D and 3D spaces for the stage by stage features retrieval processes, allows the architecture to be as flexible as the designer wants, considering that this is not a lucky fact. With the recent ten years of technological progress now we can compute and train those architectures and they have faced so many challenges for applications originating the most famous CNN architectures. This chapter presents an author position related to the artificial intelligence field and machine learning/deep learning appearance in the scientific world scene describing hastily the basis for each one and later, focusing on medical applications most of the socialized on the Annual IEEE Engineering in Medicine and Biology Society conference held in Hawaii in July 2018. While addressing the medical applications from cardiovascular to cancer diagnosis, we will briefly describe the architectures and discuss some features. Finally, we will present a contribution to the deep learning by introducing a new architecture called Convolutional Laguerre-Gauss Network with a kernel based on a spiral phase function ranging from 0 to 2π and a toroidal amplitude band-pass filter, known as the Laguerre-Gauss transform.

10.1 Introduction

In the context of the broad positioning of artificial intelligence thanks to the effects of globalization generated by the large computer companies called "GAFA": Google, Amazon, Facebook and Apple; it is imperative to take up the fundamental theoretical

O. L. Q. Montoya (✉)
Mathematical Sciences Department at Universidad EAFIT, Medellin, Colombia
e-mail: oquinte1@eafit.edu.co

J. G. Paniagua
Engineering Faculty, Instituto Tecnológico Metropolitano, Medellin, Colombia
e-mail: juanpaniagua@itm.edu.co

L. Nanni et al. (eds.), *Deep Learners and Deep Learner Descriptors for Medical Applications*, Intelligent Systems Reference Library 186,
https://doi.org/10.1007/978-3-030-42750-4_10

aspects, the variety of technical aspects, the relevance of applications in different areas and no less important, the ethical implications surrounding intelligent systems. In the framework of "*Deep Learners and Deep Learner Descriptors for Medical Applications*" this chapter aims to demystify the aspects that have idealized machine learning and artificial intelligence and discuss the elements that allow the global context to be key to the next developments that will increase the economic capacity of the countries, the research advances and increase of the level of life quality of the human beings focusing on medical applications.

Within the framework of the 2017 International Congress of the International Federation of Automatic Control (IFAC) in Toulouse, I experienced the unimaginably stunning capacity of the effect of globalization and technology by getting to know big spotlights up close of technological, technical and economic development. However, the most disruptive experience was entering a bookstore looking for what would be my reading material in French for the next trips I would undertake. The concern of the French regarding the future and how the work and the economic model are changing led the group called "*Les economistes aterrés*" to write *Changer dávenir reinventer le travail et le modele économique, 2017*. With such revealing title, the least I thought I could find was a position that determined my desire to contribute to this book.

Les Economistes aterres is a collective of economists and citizens dedicated to promoting collective reflection and the public expression of economists who do not resign themselves to the domination of neoliberal orthodoxy. As I progress in the book I find an interesting proclamation: "*Due to the GAFA, the large computer companies with out Google, Amazon, Facebook, and Apple, we live the first moments of an acceleration of technical progress unprecedented from the field of robotics and artificial intelligence. For a long time, and even today, most of the time, robots were very efficient for a particular task, they could not learn and perform simple operations for a human being. But this must change quickly thanks to deep learning, artificial intelligence and neural networks that allow machines to learn by themselves, even to be creative, suggesting that the era of intelligent robots is only a matter of time.*" Leseconomistesaterres [29].

To the previous assessment, I would add the effect of the film and television company and the mass of online services that make people dependent on technology for decision making. However, we must remember artificial intelligence is not a new area of knowledge and much less an invent of television.

Historically, automatic control and robotics allowed the increase of industrial production capacity and speed up the growth of the world's largest economies. As a concept, automatic control can bring different problems when the mission is to develop machines and tools that perform repetitive tasks that operators were previously developed and with them, the ethical questions about the number of eliminated jobs are the daily life. One premise of automation was to increase the quality of the workplace, in such a way that an operator would stop performing an uninterruptedly boring task that might endanger him, to move to supervisory and control tasks where human decisions were fundamental. The machine that only serves as an assistant and

tends to increase the speed, efficiency and effectiveness of certain jobs in jobs could not make these decisions.

The theoretical aspects and advances in the matter led the industry to make millionaire investments and the large research centers achieved advances that have the powers in their levels of economic development. However, decision-making was still part of the work of humans since the robots were not capable of performing simple tasks for a 3-month-old child like recognizing their mother's face.

The human being and his complex scheme of reasoning, learning and decision making becomes the goal that can eventually allow robotic machines to achieve the maximum of the productivity and efficiency curve. However, imitating the human brain is not as simple as imitating the work of your arms to paint a high-end car or plane. Try to imagine medical decisions during a surgery!

Composing schemes of reasoning, creation, and development of intelligence was after the Second World War (and still is) the holy grail of the theorists and practitioners because with it, would be perfect (in the human sense) the task already covered in terms of repeatability. Furthermore, they would prevent the lives of human experts trained for many hours from being risky in dangerous missions, putting high-level military systems in imbalance. Human functions such as creativity and the refinement of group decision-making schemes, pattern recognition or others, were not automated, and it was what we needed to move to the next level.

As a field, artificial intelligence was created amid the theoretical and practical developments of Control Theory when, in the 60s, engineers, mathematicians and physicists such as those formed by John Ralph Ragazzini, Lofti Zadeh and others such as neurologist Warren Mc Cullock and the logical Walter Pitts raised the theoretical foundations of fuzzy sets and neural networks seeking to develop strategies to understand how the brain performs tasks. The basis of logical reasoning under uncertainty and the principle of learning by training solidified until reaching in the 90s a boom that was the object of the scientific interests of many institutes throughout the scientific community. During that time the concepts and algorithms that are now the basis of what is now called data analytics, data and information processing and deep learning were developed.

Appearing a new concept of sets that claimed to have a better entropy than the theory of classical probability seemed scandalous and with it, a new form of logic and operators that allowed the computability of rules of knowledge revolutionized how reasoning was given to the algorithms. Linguistic, psychological, engineering and mathematical elements coalesced into a new theory. The University of California at Berkeley had the best theorists and pundits who brought fuzzy concepts to the industry and not yesterday, the world became an intelligent technological world Zadeh [71].

The academic recognition of fuzzy systems consolidated in 1978 with the creation of the scientific journal Fuzzy Sets and Systems, which compiles the advance of the theory and the application of fuzzy sets and systems. In the field of information processing, fuzzy sets are important in a grouping, data analysis and data fusion, pattern recognition, and vision. Modeling based on fuzzy rules has been combined with other techniques such as neural networks and evolutionary computation applied

to systems and control with applications in robotics, complex process control, and supervision.

Machines reasoned based on uncertainty defined in their systems of rules and supervisory systems and expert control began to was erected among the most advanced industries that required exponential improvements in their indicators and decrease in failures. Moreover, the level of uncertainty contained in the measured variables and the notions of the possibility that an element belongs to one or more groups brought with it algorithms that semi-supervised deliver patterns of behavior on an unknown set of data.

Shannon's information theory brought with it the ability to simplify and contract and get compact representations of data that came from the real world. Constructing black box models used in closed-loop control schemes and supervision were positioning in telecommunications, energy, economics and it is not too much to affirm that the result of the advances in space matter depended on the theoretical advance and the practical capacity of those who endeavored to develop it. Artificial intelligence is much more than the imaginary of the machines that dominate the world and contain humans in a matrix. The universe of the human mind, of its capacity to learn and create, are the object of previous and current research. "Learning as compression of information" is how Rissanen defined it in the framework of the Minimum Description Length principle, the maximum over which the entire area moves Rissanen [50]. But what is information compression? What is information? How do humans naturally extract it and use it to make decisions?

To review briefly the main elements of intelligent algorithms, it is necessary to present the reader with a historical/technical perspective. In the world of expert systems, which seek to imitate human expertise to either monitor or warn about certain behavior patterns of a dynamic system based on their data or rules, scientific journals have been positioned to publish articles related to all aspects of knowledge engineering, including individual methods and techniques in the acquisition and representation of language and its application in the construction of systems. Traditionally applied in software engineering and human-machine interaction, recently there has been an interest in growing markets for this type of technologies such as business, economics, market research and medical care and health [21, 35, 44].

Around the learning of machine and in particular of the Bayesian Inference it is undeniable that its effort has been to obtain algorithms that based on a priori information and some assumed distributions, offer representations that allow that the data can be re-generated again with the smallest error of the possible estimate. The challenge is to propose compact models and retrieve the parameters of the distributions from the data. On the other hand, some aspects of pattern recognition have been addressed by means of grouping algorithms that naturally mimic the ability to abstract or build a simplification of a new environment for human beings, making obvious approaches to groups that minimize the distance between its members and maximize the distance between them. Relevant applications appear in data mining and natural language processing [36].

In relation to the power of discernment of human beings in terms of classification, the developments of algorithms with linear and nonlinear Kernels have allowed

interesting mixes of statistical and probabilistic approaches and the consolidation of statistical learning that allows machines to differentiate data classes. However, the concern about the type of data on which intelligent decisions should be made has allowed the development of extraction strategies of characteristics that mostly have bases in the temporal, frequency and spatial approaches of signal processing, analysis of Fourier and the Wavelet families as a solution to the heat equation [17, 48].

So far a very elegant set of algorithms that have tried to emulate the natural development of a human that distinguishes, approximates, but does not necessarily dominate a task. A different level of approximation brings with it the possibility of imitating the sets of neuronal units working so that their connections are reinforced to specialize in the realization of this assignment. The notion of information compression has not been abandoned, but the concept of learning is reformulated on supervision and training. Good or bad raw or processed representations of the real world and clues that allow input pulses to match the systems with the signals that are obtained from it. Why not then propose a form of learning that manages to find a representation so that it is the repetition of the task who makes it an expert and not necessarily a list of rules of behavior?

Corresponding to the creation of fuzzy systems, the paradigms of neural interconnection and the models of potential architectures developed rapidly. The interpretation of the cognitive and functional processes of the brain (from evolutionary, neurological and biological perspectives) brought in with it an immense theoretical and practical advancement of the concept of learning and algorithms based on kernels influenced the establishment of neural networks that involve optimization schemes for obtaining the proper parametric set to complete the task for which they are being trained [31].

Approximations to the architecture of the neuronal tissue allowed the accomplishment of the multilayer perceptrons and the development of the learning algorithm of retro-propagation of the error (backpropagation) that is based on the solution of the problem of minimization of the instantaneous energy average of the error of all the available patterns that the network must learn [58]. New architectures with different specialized functions derived and from there, with radial base activation functions such as the Gaussians, the information is propagated through the hidden single-layer network and at that moment was the solution to the problems of exact interpolation although with computation and dimensionality difficulties. However, its use extended from finance to hybrid systems with micro-cells [4]. As for systems with high-frequency signals, the economic ones began to give problems since the amount of energy and information contained in them was not easily represented. Wavelet Networks were proposed whose architecture allows making links forward and between the input layer and the output layer directly using wavelet kernels as activation functions [34, 46].

Jang in 1993 Jang [24], questioned how to extract the best of both worlds. The ability to handle the uncertainty of fuzzy systems and the genius of training that allowed achieving greater precision in the task. The adaptable neuro-fuzzy systems ANFIS offered in two steps to minimize the problem of the selection of the shape of the categories or fuzzy sets and estimate the parameters of the neural network. Its

only problem could be the relevance of the selection of the potential entrances to the problem to be solved. Nevertheless, it remains an adequate and widely used model in the world of power systems [25].

Although ANFIS represent a solution to some kind of problems, they require well-defined patterns to achieve maximum performance. More recently and thanks to the increase in the dimensionality of the data and with it the lack of input patterns in some of the applications forced to resume concepts developed in the late 80s.

So far the reader can deduce that all the progress in the area that have led to the algorithms recognize objects and differentiate patterns, locate trends and even make purchase suggestions are the result of interdisciplinary scientific developments that seek to represent how the human being learns, infers and makes decisions. The issues that result from this can be seen from two perspectives. The first one is related to the theoretical questions about learning and the way in which these algorithms learn and simplify information, and his goal is to find a magical recipe that explains why and how to reproduce the complex interactions that give rise to the brain and that can be mathematically generalizable.

On the former perspective, we continue making efforts ranging from the study of diffusion kernels to the study of compact representations of large amounts of data in the form of graphs generated with entropy-based methods. The students of many universities of the first world and Colombian join their efforts in understanding the capacity of simplification of diverse architectures and struggle to achieve a generalizable theory.

The second perspective addresses the ethical implications surrounding the ability of machines to make suggestions for human decision making.

In this chapter, we will dedicate our examples to address the previous issues in the context of biomedical deep learning advances and present a review of recent advances on Deep Learning, transfer learning and image feature extraction using CNN, including discussing their application to other fields. Finally, a contribution for future applications will be presented, looking for the spread of the creative process for CNN rather than just the application of previous architectures that may not perform very well on a human health and medical applications.

10.2 On the Learning of Deep Learning

Deep learning took its first steps with cascading autoencoders architectures whose added value is that they are an unsupervised algorithm that applies the back-propagation algorithm to give equal outputs to the inputs. If some of the entries are correlated, then the algorithm is able to discover some of these correlations. In fact, this simple autoencoder (AE) often ends up learning a low-dimensional representation very similar to the one obtained with the principal component analysis. But although the number of hidden neurons is large, sometimes greater than the number of entries, it is possible to discover interesting structures by imposing restrictions on the network.

One of the applications that represented challenges was the processing, identification, and analysis of images. The artificial vision that was being developed required taking advantage of the capacity to simplify the information of the architectures that, like the autoencoders (AE), did, among other things, the task of extraction and selection of characteristics. New architectures were developed that exploit simplification capabilities of two-dimensional kernels with the ability of multilayer perceptrons to establish nonlinear relationships.

A posteriori, the convolutional networks were the first solutions to the problem of low error propagation (which occurs thanks to a decrease in the local gradients of the input layer neurons as the number of hidden layers increases) and to the problem of lack of patterns and need to simplify information. These networks are structured from convolution blocks, non-linearity (ReLU), pooling or subsampling and classification with a completely connected neural network as a multilayer perceptron. Its particular architecture imitates the way in which human beings construct patterns that allow us not only to recognize objects but to generalize this function.

Something like when a small human being is taught the difference between a table, an armchair and a chair, when all of them can have 4 legs; What are the relevant characteristics and not common to them that allow to differentiate them?

The primary objective of the convolution layer is to extract features from the input image, the convolution preserves the spatial relationship between the pixels by learning the characteristics of the image using small squares of the input data. The non-linearity layer allows the extracted characteristics to be distributed in a space in such a way that they are mostly differentiable. The sub-sampling reduces the dimensionality of each feature map but retains the most important information, this may be the maximum, the average, the sum, etc. Classification is done with a traditional multi-layer perceptron with softmax activation functions or support vector machines with non-linear kernels if necessary.

The most famous convolutional networks are among others: LeNet 1990, Alexnet 2012, ZF net 2013, GoogLeNet 2014, VGGNet 2014, ResNets 2015 and DenseNet 2016.

On the basis of the traditional neural network, the convolutional neural network (CNN) works as a feature extractor, adding up the convolution layer and the subsampling layer. It is the combination of the artificial neural network and backpropagation algorithm, which simplifies the complexity of the model and reduces the parameters.

The introduction of AlexNet deepened the CNNs, making the training result more accurate. It has five convolution layers and three fully-connected layers, of which conv1, conv2 and conv5 layers are connected with the max-pooling layers.

Since fully-connected layers require the fixed dimensions of the feature map, the team set the input size 820 in length on 1D-CNNs and 256×256 on 2D-CNNs. Compared with common CNN, AlexNet has the following advantages:

1. The dropout layer is used after the fully-connected layer. Randomly ignoring some neurons in the training process can alleviate the over-fitting problem
2. Max-pooling layers are used to increase the richness of features

3. The nonlinear activation function ReLU is used to speed up the forward propagation process and solve the problem of gradient explosion.

10.3 Medicine and Biology Cases

10.3.1 ECG Classification with Transfer Learning Approaches

Effective detection of arrhythmia is a critical task in the remote monitoring of electrocardiogram (ECG). The conventional ECG recognition depends on the understanding of the clinicians' experience, but the results suffer from the probability of human error due to the fatigue. To solve this problem, Wu et al. [70] proposed an ECG signal classification method based on the images to classify ECG signals into normal and abnormal beats by using two-dimensional convolutional neural networks (2D-CNNs). In their paper, they employed the AlexNet-like CNNs model.

First, contrasted the accuracy and robustness between one-dimensional ECG signal input method and two-dimensional image input method in AlexNet network. Then, to mitigate the over-fitting problem in the two-dimensional network initialized AlexNet-like network with weights trained on ImageNet,[1] to fit the training ECG images. A posteriori fine-tune of the model will further enhance the ECG classification.

The performance demonstrated on the MIT-BIH arrhythmia database shows that the Wu's method can bring about the accuracy of 98% and maintain high accuracy within Signal Noise Ratio (SNR) range from 20 to 35 dB. Their experiment shows that the 2D-CNNs initialized with AlexNet weights perform better than a one-dimensional signal method without a large-scale dataset Wu et al. [70].

10.3.2 Classification of Neurons from Extracellular Recordings via CNNs

Buccino et al. [7] reported a study with their deep learning method for classification of neurons from extracellular recordings (CNER). CNER is mainly limited to excitatory or inhibitory units based on the spike shape and firing patterns. The narrow waveforms are considered being fast-spiking inhibitory neurons and broad waveforms excitatory neurons. Their approach combines detailed biophysical modeling and powerful machine learning to classify neurons from Multi-Electrode Arrays (MEA) simulated recordings.

They used Tensorflow to train the CNNs with the following configuration: the 10 $\times$ 10 amplitude and width images are input to a 32-deep convolutional ReLU layer

[1] Typical in transfer learning can be seen in http://www.image-net.org/.

which filters the input image with 3×3 kernels with stride equal to 1. Max pooling is then applied and the image is shrunken to a 5×5 footprint.

Another 64-deep convolutional ReLU layer applies 3×3 kernels and max pooling reduces the output image features to a 3×3 size. The 3×3 features are input for a fully connected layer with 1024 artificial neurons and 2 output nodes for the binary classification and 13 for the m-type classification.

The dropout method is carried out to avoid overfitting [1], with a dropout rate of 0.7. They minimize Softmax cross entropy with the Adam optimizer during training for 5000 epochs, each time sampling 1000 observations from the dataset.

The results demonstrate that binary classification between excitatory and inhibitory types is very robust, despite the overlap between excitatory and inhibitory cells regarding spike amplitudes and widths. With the same method, authors classified 13 different cell types and established that the accuracy depends on the alignment between neurons and the MEA. Their findings are only based on simulations consequently, verification with real data is a required and important step.

No noise was included in the simulated recordings, with the rationale that sorted spikes can be cleaned by applying spike triggered averaging. I will look forward to knowing when they use electrophysiology to match it with imaging techniques to reconstruct neuron morphology, required to validate the model predictions.

10.3.3 Cardiovascular Images Analysis and Enhancement

In the study entitled "Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis"; Zreik et al. [72] presented a combination of a multi-scale convolutional neural network with an unsupervised convolutional auto-encoder. As elementary as it seems for deep learning experts, I admire their application because of its impact on the life quality of patients.

Usually, in patients with coronary artery stenoses of intermediate severity, the functional significance needs to be determined. The multi-scale approach enables the network to exploit both detailed local characteristics and contextual information for identification of patients.

Fractional flow reserve (FFR) measurement, performed during invasive coronary angiography (ICA), is most regularly used in clinical practice. Try to imagine avoiding such an invasive procedure and allows physicians an improved an accurate method for FFR measurement!

To our benefit, someone concerned about reducing the number of ICA procedures. Zreik and his team reached the automatic recognition of patients with functionally significant coronary artery stenoses, using deep learning analysis of the left ventricle (LV) myocardium in rest coronary CT angiography (CCTA). The study included consecutively acquired CCTA scans of 166 patients with FFR measurements. To identify patients with a functionally significant coronary artery stenosis, the study proposed was performed in several stages:

1. They segment the left ventricle myocardium using a multiscale convolutional neural network (CNN). The first set consists of 3 patches of 49 × 49 voxels and the second set of 3 patches of 147 × 147 voxels. The latter set of patches is down-sampled by an additional 3 × 3 max-pooling layer resulting in patches of 49 × 49 voxels, as well.

   To analyze both sets of patches, the CNN consists of two identical subnetworks. It fuses together both subnetworks in a fully connected layer, followed by a softmax layer with two units providing a classification label of the voxel at hand. Each subnetwork consists of three streams are combined together in a fully connected layer.
2. To identify the segmented LV myocardium, it is subsequently encoded using unsupervised convolutional autoencoder (CAE). The input for the CAE is an axial patch around a myocardial voxel. The encoder comprises one convolution layer followed by max-pooling and a fully connected layer with 512 units. The decoder consists of one fully connected layer, one upsampling layer followed by a convolution layer providing the reconstructed input. Thereafter, subjects are classified according to the presence of functionally significant stenosis using an SVM classifier based on the extracted and clustered encodings.

During their conference, authors declared that quantitative evaluation of LV myocardium segmentation in 20 images resulted in an average Dice coefficient of and an average mean absolute distance between the segmented and reference LV boundaries of 0.7 mm. Their results confirm that automatic analysis of the LV myocardium in a single CCTA scan collected at rest, without assessment of the anatomy of the coronary arteries, can be used to identify patients with functionally significant coronary artery stenosis. Hopefully, their discoveries will be promptly implemented and impact so many cardiology units.

On the other hand, Lessmann et al. [30] have produced an approach using convolutional neural networks (CNNs) for segmentation of calcifications in the coronary arteries, aorta and heart valves. Their solution was to apply two CNNs consecutively to first identify and label candidates (CNN1), and to finally identify true calcifications among the candidates (CNN2).

The CNN1 is therefore constructed as a purely convolutional network. Previous publications (check Lessmann et al. [30], paper references for further information) showed that spatial information is particularly important for calcium detection. To allow CNN1 to infer spatial information from the image area covered by its receptive field, its receptive field needs to be relatively large.[2] To allow for a large receptive field while keeping the number of trainable parameters low, authors rely on dilated convolutions, which are based on convolution kernels with spacing between their elements.

Their proposed CNN1 detects potential calcifications based on appearance and spatial context and furthermore determines whether the calcification is placed in the coronary arteries, the aorta, or the aortic or mitral valve. However, metal artifacts,

[2]However, CNNs with large receptive fields, such as very deep networks or networks with large convolution kernels, often suffer from overfitting due to large numbers of trainable parameters.

image noise or other high-intensity structures, such as parts of the spine in direct proximity to the aorta, can result in false positive voxel detections. Their second system called CNN2, refines the output of CNN1 by differentiating between true calcifications and false positive voxels with similar appearance and location. In contrast to CNN1, CNN2 does therefore not need to focus on the spatial context but can focus on local information and finer details. CNN2 does not use dilated convolutions, but instead non-dilated convolutions with max-pooling between convolutions. CNN2 is not purely convolutional like CNN1 as it only needs to analyze a limited number of voxels.

The CNN2 analyzes 2.5D inputs and has a receptive field of 65 $\times$ 65 pixels. Opposed to the multi-class output of CNN1, the output of CNN2 is binary as its purpose is false positive reduction and not the categorization of the detected calcifications. The technique has been assessed with diverse sets of CT scans, namely cardiac CT scans, attenuation correction CT scans acquired with a cardiac PET, chest CT scans acquired clinically and in lung cancer screening, and radiotherapy treatment planning chest CT scans. The results manifest that it yields fully automatic and accurate segmentation of calcifications. This allows judgment of CVD risk as a requested or unrequested finding in CT scans visualizing the heart [30].

Another experience I learned of, was presented by the team from Imperial College of London which showed that deep learning concepts can solve a range of tasks in cardiac Magnetic Resonance Imaging (MRI) ranging from image reconstruction to image super-resolution enhancement.

Schlemper et al. [56] considered reconstructing 2D dynamic images with Cartesian sampling using Convolutional Neural Networks (CNNs). Comparable to the formulations in compressed sensing (CS) MRI Schlemper et al. [57], authors examine the reconstruction problem as a de-aliasing problem in the image domain. It pointed out that constructing a Deep learning solution will require a deep knowledge of both image processing and machine learning.

However, reconstructing an under-sampled MR image is demanding because the images typically have a low signal-to-noise ratio, yet often high-quality reconstructions are desired for clinical applications. To deal with this issue, they used a very deep network architecture which forms a cascade of CNNs. The results show that the CNN approach is capable of producing high-quality reconstructions of 2D cardiac MR images; and the proposed method is fast enough to permit the real-time applications.

Super-resolution (SR) is another challenge. Most clinical cardiac MR images are acquired as multi-slice 2D imaging in order to reduce the length of the image acquisition and of the breath-holds. This hampers visualization and quantitative measurements as often relatively thick slices are acquired.

Oktay et al. [39] proposed a unique image super-resolution (SR) approach that is based on a CNN model. Basically, their method includes a segmentation block and super-resolution block with four convolution layers and four deconvolution layers with stalked, inverse convolutions and concatenation operations.

A stacked convolutional auto-encoder (AE) network, is trained with segmentation labels. The AE model is coupled with a predictor network to obtain a compact nonlinear representation that can be extracted from both intensity and segmentation images. The full model is named as T-L network used as a regularisation model to enforce the model predictions to follow the distribution of the learned low dimensional representations or priors.

The T-L model was trained in two stages:

1. The AE is trained separately with ground-truth segmentation masks and cross-entropy loss. The predictor model is trained to match the learned latent space by minimizing the Euclidean distance between the codes predicted by the AE and predictor. Once the loss functions for both the AE and the predictor converge, the two models are trained jointly in the second stage.
2. The encoder is updated using two separate back-propagated gradients and the two loss functions are scaled to match their range. The first gradient encourages the encoder to generate codes that could be easily extracted by the predictor; while the second gradient guarantee that a good segmentation-reconstruction can be obtained at the output of the decoder.

Results illustrate that this anatomically-constrained CNN-based Super Resolution technique produces very positive results, even with the existence of artifacts in the input data Schlemper et al. [56].

10.3.4 Nuclear Medicine Recent Applications

As mentioned in this chapter, for applications in cardiovascular medicine imaging, deep learning has outperformed the traditional machine learning and Bayesian approaches. Some examples can be found in many applications such as image restoration and super-resolution with the large dataset and high computing power graphical processing unit.

One interesting success is related to positron emission tomography (PET). Although PET is a sensitive and quantitative imaging tool that provides a functional, biochemical and molecular information via maximum likelihood reconstruction of activity and attenuation (MLAA); it suffers insufficient anatomical information, low resolution, and high noise level including crosstalk artifacts, slow convergence speed, and noisy attenuation maps (μ-maps).

New signs of progress in noise and artifact reduction in simultaneously reconstructed activity and attenuation images from only the emission PET are described in Hwang et al. [23]. These advances were carried out by training some convolutional neural networks (CNNs) that learn computed tomography (CT) derived PET attenuation maps from simultaneously reconstructed activity and attenuation data. For this procedure, three different CNN architectures were designed and trained: a convolutional autoencoder (CAE), U-net, and a hybrid of CAE and U-net. The

CNNs had to learn x-ray computed tomography (CT) derived μ-map (μ-CT) from the MLAA-generated activity distribution and μ-map (μ-MLAA).

It is time now to introduce briefly the U-net. This is a convolutional network architecture for fast and precise segmentation of images; developed by a team from the Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany[3] for fast and precise segmentation of images. Up to now, it has surpassed the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. The u-net has won the Grand Challenge for Computer-Automated Detection of Caries in Bitewing Radiography at ISBI 2015, and it has accomplished the Cell Tracking Challenge at ISBI 2015 on the two most challenging transmitted light microscopy categories, Phase contrast and DIC microscopy, by a large margin [52].

Let us go back to the artifact reduction in PET application performed by Prof Lee team.[4] They used PET/CT data of 40 patients with suspected Parkinson's disease for five-fold cross validation. For the training of CNNs, 800,000 transverse PET slices and CTs augmented from 32 patient data sets were used. The similarity to μ-CT of the CNN-generated μ-maps (μ-CAE, μ-Unet, and μ-Hybrid) and μ-MLAA was compared using Dice similarity coefficients. The CNNs provided less noise in images and more uniform μ-maps than original μ-MLAA. Moreover, the air cavities and bones were better resolved in the designed CNN outputs. In addition, these architectures were appropriate for alleviating the crosstalk trouble in the MLAA reconstruction. The hybrid network of CAE and U-net yielded the most similar μ-maps to μ-CT with Dice similarity coefficient in the whole head = 0.79 in the bone and 0.72 in air cavities, arising in only approximately 5% errors in activity and biding ratio quantification [23].

10.3.5 Neuroimaging for Brain Diseases Diagnosis

The challenges in the radiation image generation include Alzheimer and dementia diagnosis. Alzheimer's condition assessment without coregistered anatomical magnetic resonance imaging (MRI), requires a proper spatial normalization of amyloid PET images. To overcome this issue, the team of Prof. Jae Sung Lee, proposed deep learning-based self-generation of PET templates for amyloid PET spatial normalization using supervised deep neural networks. In the proposed approach, deep neural networks are trained to produce the best individually adaptive PET template.

Their conclusions, lead to progress in dealing with kidney parenchyma in computer tomography (CT). Deep learning to calculate the automatic volume of interest drawing in Glomerular filtration rate (GFR) was reported. Their work assumes that quantitative single photon emission tomography (SPECT)/CT is potentially useful

[3] https://lmb.informatik.uni-freiburg.de/index.php.

[4] From the Department of Nuclear Medicine, Seoul National University College of Medicine.

for more systematic and reliable GFR measurement than conventional planar scintigraphy. However, manual drawing of a volume of interest (VOI) on renal parenchyma in CT images is labor-intensive and time-consuming activity usually taking around 15 min per scan. Contraction and extraction operations in a 20+ layers CNN seems to be a result to be published shortly.

In recent years, some research has been reported with the premise of decoding brain areas, in which "begin" or separate certain mental states, actions or responses to an environment. Huth et al. [22], proposed by fMRI techniques perform a semantic map of the cerebral cortex, concluding that certain areas of the cerebral cortex are activated specifically by listening and thinking in certain semantic domains, such as words related to time, emotions, numbers, locations, etc. Emotion recognition or determination of emotional states will provide insigths for diagnosis of affective spectrum diseases, commonly related to dementia.

As exposed during this chapter, a recent success in deep learning is mostly relied on the convolutional neural network (CNN). It has been also successful in identifying biomarkers to support the clinical diagnosis using medical images. However, while CNN is best applicable to images since the nature of the brain network differs from the images, it is unclear that CNN is applicable to the brain network (BN). Han et al. [20] discuss this issue by comparing CNN with a simple feed-forward neural network. The authors used the brain network extracted from the diffusion-weighted MR images to investigate the applicability of CNN to the brain networks.

The construction of BN is not a straightforward process and it supposes the selection of a measurement of connectivity (depending on the signal source) providing different network configurations, not necessarily easy to define in certain diseases or as a matter of example lesions in the brain or emotional states location. The used CNN consists of 1-Dimension convolutional filter with the following features:

- Filter width of 5, 256, 64, 32, 10
- ReLu activation function for each layer
- Max-pooling of 2
- To minimize the effect of the fully-connected layers on the final performance, they used single fully-connected weight on the output with a sigmoid activation function.

To compare their network, they trained a simple feed-forward neural network with 3 layers.[5] In both models implemented in TensorFlow, they evaluated the performance using 5-fold cross-validation. But in fact, while CNN exploits local spatial topology and robustness to translational invariance, those are fairly erratic in the brain networks. For a specified network the order of nodes in the adjacency matrix affects the design of the adjacency matrix. Indeed, a bidimensional kernel can be applied for image feature extraction as in previous studies; but the adjacency matrix composition processed with a convolution filter does not yield such an expected amount of

[5] With 1000, and 200 neurons with hyper-tangent activation functions for two hidden layers, and a single output neuron with a sigmoid activation function.

information recovered. Consequently, traditional convolution cannot capture local features on the adjacency matrix. It may be the reason that CNN under-performed.

Wistfully, this research did not match the number of parameters between models since an enlarging number of parameters in CNN declined the final performance. It should be deeply explored. There are other deep learning approaches on brain networks, such as the spectral network [6], and BrainNetCNN [26], which will be studied by Prof Han's team. Though, some intuitions for reducing the troubles lead to pre-training, fine-tuning and regularization as a critical step in the Convolutional neural network architecture for this specific application.

BrainNetCNN the first CNN regressor for connectome data, is consisted of novel edge-to-edge, edge-to-node and node-to-graph convolutional filters that weight the topological locality of structural brain networks. Researchers based the composition of BrainNetCNN (for connectomes) on a typical CNN where the first section of the network has convolutional layers and the last section posses fully connected (FC) layers. A brain network described by its adjacency matrix (90×90) is the input to a BrainNetCNN model; while the output layer of the network has two nodes where each node foresees a different neurodevelopmental outcome score. The second to last layer in the network can be understood as a set of high-level features learned by the previous layers and their dimensions are $1 \times 1 \times 30$.

BrainNet behaved very well, predicting motor and cognitive scores with the highest correlations to the ground truth scores from several databases. Furthermore, it was found that, with respect to most accuracy measures, minor modifications of the core architecture (e.g., E2Enet-sml, 2E2Enet-sml) were able to surpass other models without relying on the large fully connected layers.

As an application, BrainNet researchers determined the capacity of their structure to learn multiple independent injury patterns to brain networks; by first predicting the input parameters of each instance in a realistic phantom dataset. Afterward, they tested the CNN on a set of preterm infant brain networks. And showed that the method is able to predict Bayley-III cognitive and motors cores 18 months into the future. Cognitive and motor scores predicted by BrainNetCNN had considerably higher correlations to the ground truth scores, than those predicted by other methods. One of the most promising results is that those edges that were learned by BrainNetCNN resulted to be important for each neurodevelopmental outcome, and were found to be predictive of better motor outcomes.

In Brain Reverse Engineering by Intelligent Neuroimaging (BREIN) Laboratory, Prof. Kyung Seong, Joon has been working along with his team on Artificial Intelligence for Neuroimaging Applications in Dementia. His conception of a broad spectrum of Artificial Intelligence techniques includes supervised, non-supervised and deep learning algorithms. Those devoted to diagnosis in very different stages of cognitive function related to the severity of the disease. I became engaged on the Differential diagnosis of dementia from Magnetic Resonance Images (MRI) via supervised learning; by recognizing different kinds of pathology such as Fronto Temporal Dementia (FTD) from Alzheimers Disease. And within the FTD differentiating either behavior variant FTD (bvFTD) referred to changes in behavior and personality; to the Primary Progressive Aphasia (PPA) that leads to loss of language skills. And

thereafter on, identifying two types of PPA such as Progressive Non-Fluent Aphasia (PNFA)[6] and Semantic Dementia (SD).[7]

Prof. Kyung pointed out how traditional machine learning based on labeled MRIs for features extraction via Laplace Beltrami operator, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) classification; usually require the use of the full image which means a pile of slices for hand-crafted features extraction. But likewise, declared the CNN capabilities to perform automatic features extraction and reducing the use of the entire set of slices reaching to the treatment of 1 slice of the image for human Medical Doctor expertise replication. They are working on CNN architectures via the inception of V3 in Google net.

10.3.6 Machine Learning for Cancer Imaging

Some advances on deep learning for cancer diagnosis will be mentioned as follows. It is unquestionable the relevance of cancer investigation so research will continue advancing as results continue being as successful as they are now. In 2016 we presented an application of Deep Convolutional Neural Networks (CNN) for the detection and diagnosis of breast tumors. The images used in this study were extracted from the mini-MIAS database of mammograms. The proposed system was implemented in three stages: (a) crop, rotation and resize of the original mammogram; (b) feature extraction using a pretrained CNN model (AlexNet and VGG); (c) training of a Support Vector Machine (SVM) at the classification task using the previously extracted features. In this research, the goal of the system was to distinguish between three classes of patients: those with benign, malign or without tumor.

Experiments showed that feature extraction using pretrained models provides satisfactory results, achieving a 64.52% test accuracy. This outcome could be improved via fine-tuning of the final layers or training the whole network parameters. Additionally, it is worth noting the impact of the data augmentation process and the balance of the number of examples per class on the performance of the system. The implemented system has three main advantages: (a) the mammograms are classified directly as with benign or malign tumor and without tumor, (b) it is not necessary to define a specific area in which the tumor is located and (c) apart from the mammogram, additional information must not be provided [15].

Dart et al. [12] developed a framework based on fully automated colorectal tumor segmentation from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). The technique is based on classifying the perfusion characteristics of tumors and grouping them using graphical models named “pieces of parts”. Their work has been progressing by combining the concept of principal component analysis (PCA), with the concept of dynamic 3D + time perfusion supervoxels. This way they

[6] Speech becomes hesitant and lacks grammatical accuracy.

[7] Lose ability to understand or formulate words in a spoken sentence.

obtain the dynamic signal intensity curves within tumors and surrounding regions and sub-parcellate the tumors into potentially pathologically meaningful groupings.

Image processing is an important part of the diagnosis and therapy mechanisms for Medical Doctor decision making. New in vivo imaging techniques such as multi-photon imaging applied to preclinical tumor modes, have the potential to further our understanding of the tumor microvasculature and follow it up under controlled conditions. Tumour neo-vasculature is very chaotic and almost futile to segment manually. The vasculature is known to be of key biological significance, especially in the study of cancer. Major effort has been focused on the automated measurement and analysis of vasculature in medical and pre-clinical images. In tumors, specifically, the vascular networks may be extremely irregular and the presence of the individual vessels may not conform to classical descriptions of vascular appearance.

Bates et al. [5] in their work "*Extracting 3D Vascular Structures from Microscopy Images using Convolutional Recurrent Networks*" produced a technique based on a deep learning concept known as convolutional long short-term memory units (ConvLSTM). It was employed to extract the vessel skeletons using an end-to-end optimization approach, contributing to excellent results. Authors proposed an approach to directly extract a median representation of the vessels using Convolutional Neural Networks. Then indicate that these two-dimensional centerlines can be purposely extended into 3D in anisotropic and complex microscopy images using the recently popularized Convolutional Long Short-Term Memory units (ConvLSTM). For the CNN portion of the network, they used (again!!!) a U-Net style composition with 2 convolutional layers at each pooling level. Each convolutional layer is followed by a Batch Normalization layer and then a (Leaky) Rectified Linear Unit (LReLU).

Another interesting work I heard about, was related to Lung cancer. This is one of the four major cancers in the world. Accurate diagnosing of lung cancer in the early stage plays an important role to increase the survival rate. Computed Tomography (CT) is an effective method to help the doctor detect the lung cancer. Lyu et al. [33] developed a multi-level convolutional neural network (ML-CNN) to investigate the problem of lung nodule malignancy classification. ML-CNN comprises three CNNs for extracting multi-scale features in lung nodule CT images.

There are two convolution layers followed by batch normalization (BN) and pooling layers. BN is used after the convolution operation and before the activation operation. It is used to reduce the internal covariate shift. The problem is formally known as covariate shift when the distribution of network activations changes between training and production stages.

Furthermore, authors flattened the output of the last pooling layer into a one-dimensional vector for every level and then concatenate them. The strategy can help to improve the performance of the model. The ML-CNN is applied to ternary classification of lung nodules (benign, indeterminate and malignant lung nodules). The experimental results show that our ML-CNN achieves 84.81% accuracy without any additional hand-craft pre-processing algorithm.

The state-of-the art in CNNs for Cancer is presented in Nanni et al. [38]. Authors presented an ensemble of CNNs for cancer related color datasets. The ensemble is built in a very simple way by training and comparing the performance of CNNs using

different learning rates, batch sizes, and topologies. The set of CNNs is simply combined with the sum rule. Fine tunning procedures were carried out and handcrafted descriptors were trained with a SVM. The most important finding of this work is that this simple ensemble outperforms the best stand-alone CNN. When the ensemble of CNNs is combined with other features based on handcrafted features, the final ensemble obtains state-of-theart performance on all the four tested datasets. Features extracted from these CNNs will then be used to train SVM classifiers.

10.3.7 On the Emotion Recognition Challenge

Machine learning and deep learning occurred to be wonderful tools and strategies to automate medical practice. Image processing is now a broader field and we have experienced a convergence between mathematics, computer science, and medicine. As striking as the recounted experiences are and success rates increase, we need more effort regarding human-computer interfaces.

Another field related to medical practice is the study of emotions. The area of automatic emotion recognition has been around for some time and is booming in many different fields. Define the emotions is a difficult matter, but for convenience, we consider the emotions as states, and these emotional states are the neurological answer to an external stimulus.

The study of the responses to external stimuli (emotion) is important because it allows establishing a psychological profile of the person which it is useful for the diagnosis of the psychopathology of the individual, allowing to create optimal treatment strategies [43, 59, 65]. However, the study of emotional states is a growing field, useful to improve the interaction between human and computer, developing artificial intelligence systems that can interact and react depending on the emotional state of the human user [3, 28, 54].

With an interdisciplinary team, we produced several intelligent algorithms for emotion recognition from speech, facial micro-expressions, and electroencephalogram Gómez et al. [16–18]; Chaparro et al. [11]; Uribe et al. [66]; Restrepo and Gomez [47]; Sierra-Sosa et al. [61]; Campo et al. [9].

Our prospect of applications go from children care to elderly and handicapped populations with potential extensions to design a emotionally aware smart classroom Celani et al. [10].

In Bustamante et al. [8], we mixed prosodic (energy of the signal), frequential (Mel frequency Cepstral Coefficients) and time approaches (Zero crossings, kurtosis, number of windows) to detect between happiness, anger, fear and sadness. We trained a neural network using those features mentioned which can detect those emotions in voice with a success rate of 95%. While in Mejia et al. [35], we extracted some features from audio signals using the wavelet transform and with those features and a fuzzy inference system, identifying between happiness and sadness. Obtaining a high accuracy of in detecting emotions associated with happiness and sadness using the Berlin Database [8].

In [67], we proposed a method for detecting negative emotions such as anxiety, disgust, anger or desperation in noisy speech signals. They propose some characteristics obtained from the discrete wavelet transform of the signal, and uses Gaussian Mixture models and Universal Background models as classifiers. They use a speech enhancement algorithm in order to improve the intelligibility of the signals, and it does improve the detection by 22% [49, 67]. Then, emotion classification was achieved with Gaussian Mixture Models and Universal Background Models. It uses the Berlin Database, but also uses an additional database called GVEESS (Geneva Vocal Emotion Stimulus Set).

Another step towards a feature extraction kernel was to transform audio signals using Daubechies wavelets and then extract a very big set of features from the different transforms. Then, performed t-tests by pairs of emotions to reduce the number of characteristics. With this reduced set a Neural Network with 50 neurons and 1 hidden layer is trained using the hyperbolic tangent sigmoid as transfer function. The method proposed had a 90.96% accuracy rate over the Berlin Database [9].

Exploring traditional machine learning strategies, in Santiago et al. [55] we extracted some 12 features: entropy based features, frequency based features, energy related features, dynamic features and paralinguistic features from the wavelet-preprocessed signal. Using Linear Discriminant Analysis and Decision Trees in order to classify the data, using a one-vs-all approach. An error of 34% in classifying the emotions was obtained. Wavelet kernels, feature selection procedures and machine learning in EEG processing for emotion recognition can be found in Gómez et al. [16–18].

Emotional features search led us to image processing bi dimensional kernels, consequently we compute the spectograms of audio signals which are in turn converted into images in grey scale. We applied a 2-D Fourier transform to this image. The spectral densities of this images are then estimated, and an average spectral density is figured for each emotion. From these images they can qualitatively assert which emotion is represented by a new audio signal by looking at its spectral density (?). Posing the feature extraction problem as a two dimensional signal represented a step on the deep learning approach for emotion recognition in speech. But audio signals are not the unique interest of our group.

Psychologists have proposed that certain microexpressions of the face correspond to certain emotions. Facial emotion detection then aims to automatically recognise different emotions based on those microexpressions. The idea of facial emotion detection is to mark specific facial features (eyes, mouth, eyebrows and nose), and based on this marks, detect the emotion of the person being studied. It is well known that Ekman and Friesen [14] proposed the Facial Action Coding System (FACS), which label emotions based on the position of facial features. There are many database used for emotion detection in faces, but one of the most complete is the Cohn-Kanade extended (CK+) database [32].

As first step, we proposed an algorithm for feature extraction on faces, based on FACS [48]. An example of Anger emotion in greyscale is depicted in Fig. 10.1. The algorithm works in the following manner: First, they read the image and preprocess it (lighting correction). Then the ViolaJones Viola and Jones [68] algorithm is applied, which is an algorithm used for object detection in images (in this case the object we

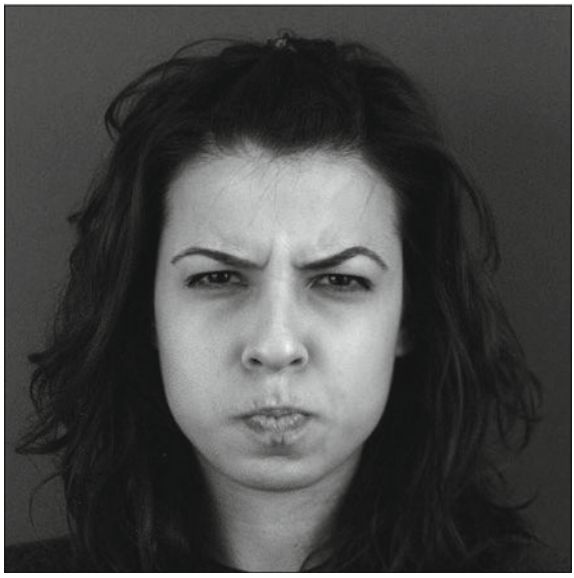


Fig. 10.1 Anger emotion in grayscale, typical image in a database

try to detect are certain facial landmarks). Then, all images are resized to 512 × 512 pixels, in order to standardise the measurements of the features.

Then the image is grey scaled and Viola-Jones is applied to detect the nose and a mark is established. Afterwards, Viola-Jones is applied to detect the eyes and an adaptive threshold method is applied in order to binarise the image (the parameters of the adaptive threshold are computed using genetic algorithms because they are faster than trying out all possible combinations of parameters), then the relevant Hough circles are computed to establish the eyes middle point and with those points you can establish the canon proportions of the face, with which you would be able to extract different features of the face.

An improvement on those algorithms was the basis for a development. With the features extracted using Rincon-Montoya et al. [48] algorithms and a new image processing kernel applied, we feed a Neural Network which can classify emotional states. As result, an overall accuracy of 90.7% was reported using the CK+ database [47]. In order to automate the features extraction and accelerate the process of online computation, next step is to fuse the bi-dimensional kernels with a convolutional neural network architecture. The research proposal will include not only the presented results but also a recent developed spatial filter with potential in medicine and biology field.

In Table 10.2 is presented the summary of datasets and their main features.

In Chaparro et al. [11] we proposed a data fusion technique in order to enhance the emotion recognition strategies. In our application we managed to put into the same architecture the features from early and late stages with machine and deep learning strategies.

10.3.8 Convolutional Laguerre Gauss Network

A really good lesson from BrainNEt Han et al. [20] team was to be inspired by the nature of the problem and not necessarily train a bunch of patterns. Our aim is to figure it our how to properly learn the features of the data and use our knowledge to design new architectures.

Several techniques have been proposed to reduce the artifacts occurrence in imaging. Derivative operators as Laplacian are the most frequently used. It is well-known that Laplacian operator boosts the high frequencies relative to the low frequencies. Therefore, the Laplacian is often used to dampen low-frequency artifacts when the background medium contains sharp wave velocity contrasts Rocha et al. [51]. Based on this, Paniagua and Sierra-Sosa [42] decided to propose the use of another method to reduce or eliminate the low-frequency artifacts and does not introduce other uncertainties in the scalar field produced in subsurface images. Paniagua's technique is based on a spiral phase function ranging from 0 to 2π and a toroidal amplitude bandpass filter, known as Laguerre-Gauss transform.

Through numerical experiments we presented the application of this particular integral transform on three synthetic data sets. In addition, we presented a comparative spectral study of images obtained by the zero-lag cross-correlation imaging condition, the Laplacian operator and the Laguerre-Gauss transform, showing their spatial frequency features. We also presented evidences not only with simulated spatial noisy velocity fields but also by comparison with the velocity field gradients of the dataset that this method improves the Reverse Time Migration (RTM)[8] scalar fields by reducing the artifacts and notably enhance the reflective events [40–42]. Examples can be seen in Figs. 10.2, 10.3 and 10.4 where the absolute value, imaginary and real fields obtained with our kernel, are presented.

Laguerre Gauss transform kernel uses a pure-phase function with a vortex structure in spatial frequency domain, defined as $B(f_x,\ f_z) = \tan^{-1}\left(\frac{f_z}{f_x}\right)$. The particular property from this spiral phase function is that is composed by a heavy-side function with a π gap when crossing the origin in every angular direction. In the amplitude, the kernel includes a gaussian toroidal geometry.

The Laguerre-Gauss transform allows to realize an isotropic radial Hilbert transform without resolution loss [19]. In addition to the advantage of spatial isotropy common to the Riesz transform stemming from the spiral phase function with the unique property of a signum function along any section through the origin, the Laguerre-Gauss transform has the favorable characteristics to automatically exclude any DC component [69].

An effect associated with the application of the Laguerre-Gauss transform is the phase and amplitude changes in the final image. Changes in amplitude are associated with the topological characteristics of the pseudo complex field and will be analyzed

[8]There is a direct relation on the physics for RTM and medical imaging check the work of Wang et al. 2016 in IEEE Transactions on Medical Imaging, vol. 35.

Fig. 10.2 Lena image processed by Laguerre Gauss convolutional kernel, absolute value

Fig. 10.3 Lena image processed by Laguerre Gauss convolutional kernel, imaginary value

in future works to find its relation with attributes such as amplitude and reflectivity of wave signals.

Properties of the Laguerre-Gauss filter allows feature extraction and artifacts removal, perfect for enhancement of the previous algorithms of cancer detection [15] and emotion recognition as in Sect. 10.3.7. Consequently, our research group next steps are to use it as convolution layer within a CNN with transfer learning stages looking for our Convolutional Laguerre Gauss Network (CLG-Net) best performance. Data base image of anger was processed with our kernel. The absolute value, imaginary and real values of the fields can be seen in Figs. 10.5, 10.6 and 10.7 respectively.

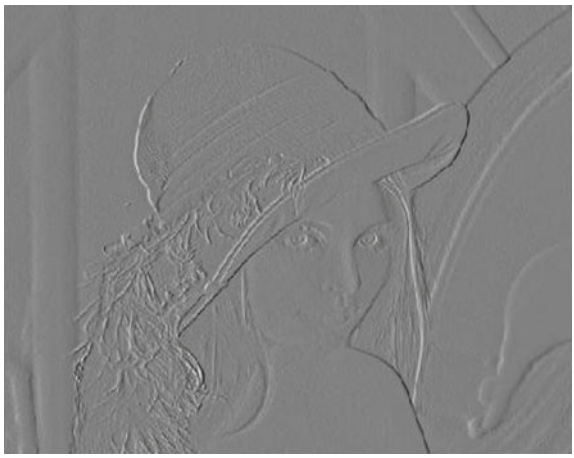


Fig. 10.4 Lena image processed by Laguerre Gauss convolutional kernel, real value

Fig. 10.5 Anger emotion processed via Laguerre Gauss kernel absolute value

10.4 Ethical and Practical Concerns

On the latter issue I pointed at the end of the introductory section, I must admit that truly we are in the era in which administration of data and information makes vulnerable the human beings.

And their use, even more, sensitive and susceptible to being diverted towards trends and tastes. This is how in March of 2018 one of our GAFA was embroiled in a scandal of misuse of data and information; or a deviation with purposes different from those of our giant.

Fig. 10.6 Anger emotion processed via Laguerre Gauss kernel imaginary value

Fig. 10.7 Anger emotion processed via Laguerre Gauss kernel real value

But under the assumption of adequate handling of the data (with which the algorithms learn); what should be the correct use or approach of the new theoretical and practical developments of artificial intelligence and machine learning?

A well know paper entitled "Scalable and accurate deep learning with electronic health records" from Rajkomar et al. [45] introduced a Predictive modeling with electronic health record (EHR) data. It is expected to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data. This is a labor-intensive process that discards the vast majority of information in each

patient's record. The team from Google Inc, Mountain View, University of California, San Francisco, University of Chicago Medicine and Stanford University developed a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. They described that deep learning methods using this representation are qualified of accurately anticipating multiple medical events from multiple centers without site-specific data adjustment. Also validated their method using anonymized EHR data from two of US academic medical centers.

In the sequential format proposed, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical observations. Deep learning models achieved high accuracy for tasks such as predicting in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and all of a patient's final discharge diagnoses. These models outperformed traditional, clinically-used predictive models in all cases able and accurate deep learning with electronic health records" from Rajkomar et al. [45]. Their approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, I could not find any detail about the architecture. The Nature Digital Medicine Journal paper contribution is twofold. As they indicate:

- Reporting a generic data processing pipeline that can take raw EHR data as input, and produce FHIR outputs without manual feature harmonization. (This makes relatively easy to deploy the system to a new hospital.)
- Secondly, based on data from two academic hospitals with a general patient population (not restricted to ICU), demonstrating the effectiveness of deep learning models in a wide variety of predictive problems and settings (e.g., multiple prediction timing) Rajkomar et al. [45]. No comparison was provided.

Is it true that a machine can perform a prognosis about if we are going to leave the hospital alive or dead? There is still no proper answer. Another issue will concern the security awareness of the information systems because medical data restrictions will lead to smart solutions. Also, the hardware industry (such as Intel) must face some challenges related to long product life, inconsistent update mechanism, unencrypted data transmission, closed systems, not known behavior, maintenance of devices and infrastructure.

Regarding the practical aspects of the development of the area, I will quote below what I believe should be the future of artificial intelligence worldwide, and that in Colombia we have made concrete and decisive progress.

Despite the technical and theoretical explosion in the mathematical sciences, the field of Human, Robot, Human Machine, Human-Computer interfaces continues to unite interdisciplinary efforts whose central axis is the human being; and in which neuropsychologists, geneticists, bioengineers, and physicists participate.

Developments in the recognition of emotions in physiological signals such as facial microexpressions, voice, electroencephalograms, data fusion with electrocardiogram signals, temperature, and body sweating are under developing in our country and globally continue to be a challenge that moves the field of intelligence artificial and machine learning and in which teams from all over the world take part. Their

findings are published in the top medicine, mathematics and engineering journals. In Figs. 10.8, 10.9 and 10.10 can be seen our preliminar results of the processing kernel proposed Convolutional Laguerre Gauss Network in real environment.

All the areas of artificial intelligence and machine learning continue in constant theoretical and practical development, applications that should in principle support human beings are still carried out. My favorite focuses are related to the increase

Fig. 10.8 Sadness expressed in real environment processed by Laguerre Gauss Kernel

Fig. 10.9 Anger expressed in real environment processed by Laguerre Gauss Kernel

Fig. 10.10 Surprise expressed in real environment processed by Laguerre Gauss Kernel

in the quality of life of patients with disabilities, early diagnosis of diseases of the affective spectrum and irregular behaviors such as child abuse, and support for the improvement of educational processes [2, 53]. Cancer and imaging implications were widely discussed in this chapter.

Nevertheless, I will keep asking me the question about how will be the best way to understand why those networks learn and how can we improve their learning capability not only by training but also by design?

Sridhar Mahadevan, Fellow of AAAI said: "… *Ultimately, the solution to challenging AI problems will have to be based on solid advances in engineering and math and physics, and I for one am confident that eventually deep learning will become transformed into a solid scientific field. That day, however, is not here yet*…"

10.5 Conclusions

It is unquestionable that thanks to the phenomenon of globalization and the positioning of the Internet, artificial intelligence covers a prevailing position in modern judgment. If not for the regular use of the GAFA technological giants, we could not be bewildered by appliances that can recognize a face or recommend to buy the shoes that satisfy all the characteristics we prefer. Much less access in a matter of milliseconds to the sources and preferences we have shown to share between contacts. The concern of many around algorithms implanted in intelligent machines is no longer necessary when human beings have diminished, instead of increasing, their potential. Some people leave their devices and algorithms deciding to choose the path to go

Table 10.1 Architectures referenced in this chapter for deep learning

Main challenges		
Challenge name	Winner	Year
Hand-written numbers on checks	Lenet	1990
ImageNet ILSVRC challenge	Alexnet	2012
ILSVRC challenge	ZFnet	2013
ImageNet ILSVRC challenge	Google Net	2014
ILSVRC challenge	VGG Net	2014
ILSVRC challenge	Res Net	2015
CIFAR-10, CIFAR-100, SVHN, and ImageNet	Dense net	2016
ISBI challenge	U-net	2015
ABIDE	Brain-net	2016

from one place to another, having taken unnecessary directions to achieve the final goal.

However, there are still possibilities of developing systems that are more intelligent than humans when there are others concerned with developing new and better theoretical bodies. Those allow someone to achieve algorithms that on the right machines can increase the quality of life within the planet or beyond.

The medical and biological fields are one of the most promising fields of progress due to the use of the remarkable world of deep learning and automatic features extraction. The DL applications will help to provide and control treatments that will improve quality of life for humanity. Improving the therapeutic practice and helping physicians and MD to provide an accurate and early diagnosis. Tables 10.1 and 10.2 provide a summary of the mentioned architectures and datasets for Deep learners. There is no ambiguity that a machine will never replace an MD expert, but machine intelligence will benefit and its aimed for human decision making.

Finally, developing solutions is not just training a preconceived convolutional/deep network, neither machine learning is just a matter of a toolbox. It requires interdisciplinary work and mathematical background to imagine and formalize a proper feature extractor closest to the human perception of the world.

Table 10.2 Available datasets referenced in this chapter

Main characteristics for medical databases and some comments for helping the reader			
Database name	Authors	Year	Comments
CK++	Lucey et al. [32]	2010	Microfacial expressions
HCI	Soleymani et al. [62]	2012	Multimodal for emotions
MIT-BIH	Moody and Mark [37]	2001	Arrhythmia database
MEA	Buccino et al. [7]	2018	Multi electrode array
CCTA	Zreik et al. [72]	2018	150 images[a]
MRI	Schlemper et al. [56], Scott et al. [60]	2018	DT-CMR
Grand challenge ISBI	ISBI	2015	Int. Sym. on Bio. Imag.
PET CTI	Team [64]	2011	1744 CT scans
ABIDE	Di Martino et al. [13], Khosla et al. [27]	2018	Autism
MINIMIAS DB	Suckling et al. [63]	2015	Breast Cancer
EHR DB	Rajkomar et al. [45]	2018	Clinical measurements
Image-net	imagenet.org	2012-x	Images[b]

[a] https://www.science.gov/topicpages/t/tomographic+angiography+ccta
[b] http://cocodataset.org

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D.G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, A., Vanhoucke, V., Vasudevan, V., Viégas, F.B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. CoRR (2016). arXiv:1603.04467
2. Akputu, O., Seng, K., Lee, Y.: Affect recognition for web 2.0 intelligent e-tutoring systems: exploration of students' emotional feedback, pp. 188–215 (2015)
3. Arnold, M.G.: Combining conscious and unconscious knowledge within human-machine-interfaces to foster sustainability with decision-making concerning production processes. J. Clean. Prod. **179**, 581–592 (2018)
4. Baghaee, H.R., Mirsalim, M., Gharehpetian, G.B.: Multi-objective optimal power management and sizing of a reliable wind/pv microgrid with hydrogen energy storage using mopso. J. Intell. Fuzzy Syst. **32**(3), 1753–1773 (2017)
5. Bates, R., Irving, B., Markelc, B., Kaeppler, J., Muschel, R., Grau, V., Schnabel, J.A.: Extracting 3d vascular structures from microscopy images using convolutional recurrent networks (2017)
6. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. CoRR (2013). arXiv:1312.6203
7. Buccino, A.P., Ness, T.V., Einevoll, G.T., Cauwenberghs, G., Fyhn, M.: A deep learning approach for the classification of neuronal cell types, pp. 1–4 (2017)
8. Bustamante, P.A., Celani, N.M.L., Perez, M.E., Montoya, O.L.Q.: Recognition and regionalization of emotions in the arousal-valence plane. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), vol. 2015, Novem, pp. 6042–6045. IEEE (2015)

9. Campo, D., Quintero, O.L., Bastidas, M.: Multiresolution analysis (discrete wavelet transform) through Daubechies family for emotion recognition in speech. J. Phys: Conf. Ser. **705**, 012034 (2016)
10. Celani, N.L., Ponce, S., Quintero, O.L., Vargas-Bonilla, F.: Improving quality of life: home care for chronically ill and elderly people. In: Caregiving and Home Care. InTech (2018)
11. Chaparro, V., Gomez, A., Salgado, A., Quintero, O.L., Lopez, N., Villa, L.F.: Emotion recognition from EEG and facial expressions: a multimodal approach. In: IEEE Engineering in Medicine and Biology Society (EMBS) (2018)
12. Dart, R.J., Vantourout, P., Laing, A., Digby-Bell, J.L., Powell, N., Irving, P.M., Hayday, A.: Tu1811-colonic gamma delta T cells respond innately to Nkg2D ligands and are grossly dysregulated in active inflammatory bowel disease. Gastroenterology **154**(6), S-1026 (2018)
13. Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., et al.: Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. Scientific data **4**, 170010 (2017)
14. Ekman, P., Friesen, W.V.: Measuring facial movement. Environ. Psychol. Nonverbal Behav. **1**(1), 56–75 (1976)
15. Gallego-Posada, J.D., Montoya-Zapata, D.A., Quintero-Montoya, O.L., Montoya-Zapa, D.A.: Detection and diagnosis of breast tumors using deep convolutional neural networks. In: Conference Proceedings of XVII Latin American Conference in Automatic Control, p. 17 (2016)
16. Gómez, A., Quintero, L., López, N., Castro, J.: An approach to emotion recognition in single-channel EEG signals: a mother child interaction. J. Phys. Conf. Ser. **705**(1), 012051 (2016)
17. Gómez, A., Quintero, L., López, N., Castro, J., Villa, L., Mejía, G.: Emotion recognition in single-channel EEG signals using stationary wavelet transform. In: IFMBE Proceedings, Claib (2016)
18. Gómez, A., Quintero, L., López, N., Castro, J., Villa, L., Mejía, G.: An approach to emotion recognition in single-channel EEG signals using stationary wavelet transform. In: IFMBE Proceedings, Claib, pp. 654–657 (2017)
19. Gou, Y., Han, Y., Xu, J.: Radial Hilbert transform with Laguerre-Gaussian spatial filters. Opt. Lett. **31**, 1394–1396 (2006)
20. Han, Y., Yoo, J., Kim, H.H., Shin, H.J., Sung, K., Ye, J.C.: Deep learning with domain adaptation for accelerated projection-reconstruction MR. Magn. Reson. Med. **80**(3), 1189–1205 (2018)
21. Hurtado Moreno, L., Quintero, O.L., García Rendón, J.: Estimating the spot market price bid in colombian electricity market by using artificial intelligence. Rev. Metod. Cuantitativos Para Econ. Empres. **18**(1) (2014)
22. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532**(7600), 453 (2016)
23. Hwang, D., Kim, K.Y., Kang, S.K., Seo, S., Paeng, J.C., Lee, D.S., Lee, J.S.: Improving accuracy of simultaneously reconstructed activity and attenuation maps using deep learning. J. Nucl. Med. Off. Publ. Soc. Nucl. Med. (2018)
24. Jang, J.-S.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. **23**(3), 665–685 (1993)
25. Karaboga, D., Kaya, E.: Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey. Artif. Intell. Rev. (2018)
26. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**(October), 1038–1049 (2017)
27. Khosla, M., Jamison, K., Kuceyeski, A., Sabuncu, M.: 3d convolutional neural networks for classification of functional connectomes (2018). arXiv:1806.04209
28. Kreps, G.L., Neuhauser, L.: Artificial intelligence and immediacy: designing health communication to personally engage consumers and providers. Patient Educ. Couns. **92**(2), 205–210 (2013)
29. Leseconomistesaterres: Changer davenir reinventing travail et le modele economique, p. 24 (2017)

30. Lessmann, N., van Ginneken, B., Zreik, M., de Jong, P.A., de Vos, B.D., Viergever, M.A., Išgum, I.: Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. Comput. Vis. Pattern Recognit. **37**(2), 615–625 (2017)
31. Linnainmaa, S.: Taylor expansion of the accumulated rounding error. BIT Numer. Math. **16**(2), 146–160 (1976)
32. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE (2010)
33. Lyu, J., Ling, S.H., Member, S.: Using Multi-level Convolutional Neural Network for Classification of Lung Nodules on CT Images, pp. 686–689 (2018)
34. Masood, Z., Majeed, K., Samar, R., Raja, M.A.Z.: Design of Mexican hat wavelet neural networks for solving Bratu type nonlinear systems. Neurocomputing **221**, 1–14 (2017)
35. Mejia, G., Campo, D., Quintero, L.: Sistema de Inferencia Basado en Lógica Difusa para la Identificación de Felicidad y Tristeza en Señales de Audio. (October), 0–4 (2014)
36. Montoya, O.L.Q., Villa, L.F., Mun˜oz, S., Arenas, A.C.R., Bastidas, M.: Information retrieval on documents methodology based on entropy filtering methodologies. Int. J. Bus. Intell. Data Min. **10**(3), 280–296 (2015)
37. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. **20**(3), 45–50 (2001)
38. Nanni, L., Ghidoni, S., Brahnam, S.: Ensemble of convolutional neural networks for bioimage classification. Appl. Comput. Inform. (2018)
39. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2018)
40. Paniagua, J., Quintero, O.: Attenuation of reverse time migration artifacts using Laguerre-Gauss filtering. In: 79th EAGE Conference and Exhibition 2017, Paris. EAGE (2017)
41. Paniagua, J.G., Quintero, O.L.: The use of Laguerre-Gauss transform in 2D reverse time migration imaging. In: 15th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio de Janeiro, Brazil, 31 July–3 August 2017, pp. 1284–1289. Brazilian Geophysical Society (2017)
42. Paniagua, J.G., Sierra-Sosa, D.: Laguerre Gaussian filters in reverse time migration image reconstruction. VII Simpósio Brasileiro de Geofísica **VII**, 6 (2016)
43. Pessoa, L.: On the relationship between emotion and cognition. Nat. Rev. Neurosci. **9**(2), 148–158 (2008)
44. Quintero, O.L., Jaramillo P.F., Bastidas, M.: Modeling perspective for the relevant market of voice services: mobile to mobile. Maskana **6**(2), 187–201 (2015)
45. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al.: Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. **1**(1), 18 (2018)
46. Rakhshandehroo, G., Akbari, H., Afshari Igder, M., Ostadzadeh, E.: Long-term groundwater-level forecasting in shallow and deep wells using wavelet neural networks trained by an improved harmony search algorithm. J. Hydrol. Eng. **23**(2), 04017058 (2017)
47. Restrepo, D., Gomez, A.: Short research advanced project: development of strategies for automatic facial feature extraction and emotion recognition. In: 2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC), pp. 1–6. IEEE (2017)
48. Rincon-Montoya, S., Gonzales-Restrepo, C., Sierra-sosa, D., Restrepo-Gómez, R., Quintero, L.: Evaluation and development of strategies for facial features extraction for emotion detection by software. (December), 1–7 (2015)
49. Ríos-Sánchez, B., Arriaga-Gómez, M.F., Guerra-Casanova, J., de Santos-Sierra, D., de Mendizábal-Vázquez, I., Bailador, G., Sánchez-Ávila, C.: gb2sμMOD: a MUltiMODal biometric video database using visible and IR light. Inf. Fusion **32**, 64–79 (2016)
50. Rissanen, J.: Minimum Description Length Principle. Springer (2010)

51. Rocha, D., Sava, P., Guitton, A.: 3d acoustic least-squares reverse time migration using the energy norm. Geophysics **83**(3), S261–S270 (2018)
52. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, pp. 1–8 (2015)
53. Rudovic, O., Lee, J., Dai, M., Schuller, B., Picard, R.: Personalized machine learning for robot perception of affect and engagement in autism therapy (2018). arXiv:1802.01186
54. Rus, S., Joshi, D., Braun, A., Kuijper, A.: The emotive couch-learning emotions by capacitively sensed. Procedia Comput. Sci. **130**, 263–270 (2018)
55. Santiago, R.C., Manuela, B.O., Quintero, M., Lucía, O.: Order dependent one-vs-all tree based binary classification scheme for multiclass automatic speech emotion recognition. (October 2014), 1486–1490 (2014)
56. Schlemper, J., Oktay, O., Rueckert, D.: Deep learning for image reconstruction and super-resolution: applications in cardiac MR imaging, p. 7 (2018)
57. Schlemper, J., Yang, G., Ferreira, P., Scott, A., McGill, L.-A., Khalique, Z., Gorodezky, M., Roehl, M., Keegan, J., Pennell, D., et al.: Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac MRI (2018). arXiv:1805.12064
58. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
59. Schutter, D.J., van Honk, J.: Extending the global workspace theory to emotion: phenomenality without access. Conscious. Cogn. **13**(3), 539–549 (2004)
60. Scott, A.D., Nielles-Vallespin, S., Ferreira, P.F., Khalique, Z., Gatehouse, P.D., Kilner, P., Pennell, D.J., Firmin, D.N.: An in-vivo comparison of stimulated-echo and motion compensated spin-echo sequences for 3 T diffusion tensor cardiovascular magnetic resonance at multiple cardiac phases. J. Cardiovasc. Magn. Reson. **20**(1), 1 (2018)
61. Sierra-Sosa, D., Bastidas, M., Ortiz P., D., Quintero, O.: Double fourier analysis for emotion identification in voiced speech. J. Phys. Conf. Ser. **705**(October 2015), 012035 (2016)
62. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**(1), 42–55 (2012)
63. Suckling, J., et al.: The mini-MIAS database of mammograms. In: Society TMIA (ed.) Digital Mammography Database ver, 1 (2015)
64. Team, N.L.S.T.R.: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**(5), 395–409 (2011)
65. Tsuchiya, N., Adolphs, R.: Emotion and consciousness. Trends Cogn. Sci. **11**(4), 158–167 (2007)
66. Uribe, A., Gomez, A., Bastidas, M., Quintero, O.L., Campo, D.: A novel emotion recognition technique from voiced-speech. In: 2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC), pp. 1–4. IEEE (2017)
67. Vásquez-Correa, J.C., Arias-Vergara, T., Orozco-Arroyave, J.R., Vargas-Bonilla, J., Arias-Londoño, J.D., Nöth, E.: Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
68. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–I. IEEE (2001)
69. Wang, Z., Ding, H., Lu, G., Bi, X.: Reverse-time migration based optical imaging. IEEE Trans. Med. Imaging **35**(1), 273–281 (2016)
70. Wu, Y., Yang, F., Liu, Y., Zha, X., Yuan, S.: A comparison of 1-D and 2-D deep convolutional neural networks in ECG classification, pp. 324–327 (2018)
71. Zadeh, L.A.: From search engines to question-answering systems the need for new tools. In: The 12th IEEE International Conference on Fuzzy Systems, FUZZ'03, vol. 2, pp. 1107–1109 (2003)
72. Zreik, M., Lessmann, N., van Hamersvelt, R.W., Wolterink, J.M., Voskuil, M., Viergever, M.A., Leiner, T., Išgum, I.: Deep learning analysis of the myocardium in coronary ct angiography for identification of patients with functionally significant coronary artery stenosis. Med. Image Anal. **44**, 72–85 (2018)