Imperial College of Science, Technology and Medicine
Department of Computing

Machine Learning in Medical Image Analysis

Liang Chen

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy

of

Imperial College London

September 2018

Acknowledgements

I would like to thank my supervisors Prof. Daniel Rueckert and Dr. Paul Bentley for their invaluable supervision through my PhD study and research. It has been my great honour and fortune to study and research under Daniel's supervision since my master year. I cannot complete my master and PhD without Daniel's insightful guidance and kind encouragement. I am also grateful to Paul for his advice and patience for teaching me clinical knowledge, motivating me with ideas, and answering my questions.

I would also like to thank the BioMedIA group, the department of computing, the department of medicine, and the college for creating an enjoyable environment. When I came to Imperial in 2012, I met many friends, colleagues, and external collaborators who helped me with my study, research, and life.

Finally, my sincere gratitude goes to my beloved family for their endless love and unconditional support. Particularly, my parents provided me with the great opportunity of studying at Imperial; and my wife always supports and encourages me with her heart.

Declaration

I declare that the work presented in this thesis in my own, unless specifically acknowledged.

Liang Chen

Abstract

Machine learning is playing a pivotal role in medical image analysis. Many algorithms based on machine learning have been applied in medical imaging to solve classification, detection, and segmentation problems. Particularly, with the wide application of deep learning approaches, the performance of medical image analysis has been significantly improved. In this thesis, we investigate machine learning methods for two key challenges in medical image analysis: The first one is segmentation of medical images. The second one is learning with weak supervision in the context of medical imaging.

The first main contribution of the thesis is a series of novel approaches for image segmentation. First, we propose a framework based on multi-scale image patches and random forests to segment small-vessel disease (SVD) lesions on computed tomography (CT) images. This framework is validated in terms of spatial similarity, estimated lesion volumes, visual score ratings and was compared with human experts. The results showed that the proposed framework performs as well as human experts. Second, we propose a generic convolutional neural network (CNN) architecture called the DRINet for medical image segmentation. The DRINet approach is robust in three different types of segmentation tasks, which are multi-class cerebrospinal fluid (CSF) segmentation on brain CT images, multi-organ segmentation on abdomen CT images, and multi-class tumour segmentation on brain magnetic resonance (MR) images. Finally, we propose a CNN-based framework to segment acute ischemic lesions on diffusion weighted (DW)-MR images, where the lesions are highly variable in terms of position, shape, and size. Promising results were achieved on a large clinical dataset.

The second main contribution of the thesis is two novel strategies for learning with weak supervision. First, we propose a novel strategy called context restoration to make use of the images without annotations. The context restoration strategy is a proxy learning process based on the CNN, which extracts semantic features from images without using annotations. It was validated on classification, localization, and segmentation problems and was superior to existing strategies. Second, we propose a patch-based framework using multi-instance learning to distinguish normal and abnormal SVD on CT images, where there are only coarse-grained labels available. Our framework was observed to work better than classic methods and clinical practice.

Acronyms

κw linear weighted-kappa scores. 41, 44, 47

AD Alzheimer's Disease. 122, 123, 128

ADC apparent diffusion coefficient. 74

ASPP atrous spatial pyramid pooling. 26, 78, 88

BN batch normalization. xxiii, 54, 56, 57, 59

BP backpropagation. 20, 21, 84

BraTS brain tumour segmentation. 52, 58, 69, 72, 76, 77

CI confidence interval. 42

CNN convolutional neural network. vii, xiv, xvii, xviii, xxiv, 5, 9–12, 20–22, 24–26, 33, 50, 51, 53, 61–63, 66, 67, 69, 70, 72, 75–80, 82, 86–90, 93, 94, 96, 132–134

CRF conditional random field. 26, 53, 72, 77, 88, 89, 96

CSF cerebrospinal fluid. vii, xiii, xvii, xxiv, 8, 10, 52, 58–61, 63–65, 72, 132, 135

CT computed tomography. vii, xii, xiii, xv, xxi–xxiii, xxvi, 1–3, 5, 6, 8–12, 33–38, 40–49, 52, 59–61, 65, 66, 72, 78, 121–124, 126–133

DNN deep neural network. 30, 31, 76, 134

DR detection rate. 10, 85, 86, 89, 95, 96

DW diffusion weighted. vii, 8, 10, 75, 94, 132

DWI diffusion weighted imaging. xiii, xxi, xxiv, xxv, 2, 3, 8, 10, 73–76, 78, 80, 82–88, 90, 92–94, 96, 98

FCN fully convolutional network. xxii, xxiii, 25, 53–55, 57, 61–63, 66, 69, 71, 78, 79, 81, 88, 89, 96, 132

FLAIR fluid-attenuated inversion recovery. xxi, xxiii, 3, 8, 34, 36, 38, 39, 42, 45, 46, 68, 74

FN false negative. 69, 70, 85, 88, 90, 92, 94, 129, 132

FP false positive. xviii, xxiv, xxv, 69, 70, 72, 75, 76, 78, 81, 83–86, 89–94, 96, 129, 132

GM gray matter. 121

GMM Gaussian mixture model. 29

HU Hounsfield unit. 2, 61

ICH intracranial hemorrhagic. 8, 49, 130, 135

ICL Imperial College London. xxii, 36, 37, 60, 86

IQR interquartile range. 37, 41–43

ISLES ischemic stroke lesion segmentation. 76, 77

IST-3 Thrid International Stroke Trial. xvii, 36, 37, 41

LSVRC large scale visual recognition challenge. 21, 22

MIL multi-instance learning. 11, 12, 122–125, 129

ML machine learning. 13

MR magnetic resonance. vii, xxi, 3, 6, 8, 10, 11, 35, 36, 38, 42, 52, 76–78, 122, 124, 129, 132, 133

MRI magnetic resonance imaging. xxiii, 1–3, 8, 9, 34–38, 41, 42, 44, 45, 48, 59, 68, 69, 74, 87, 133

NIHSS National Institutes of Health stroke score. 61

PEIS patch-based evaluation of imaging similarity. 41

PET positron emission tomography. 4

PSPNet pyramid scene parsing network. 53

ReLU rectified linear unit. xxiii, 21, 54, 56, 57

RF random forest. xix, 11, 17, 18, 31, 38, 39, 77, 121, 128, 129

ROC receiver operating characteristic. xxiii, 48, 49

ROI region of interest. xxvii, 38, 39, 122, 124, 127–129, 133

SPECT single-photon emission computed tomography. 4

std standard deviation. xviii, 41, 87, 97, 128

SVD small-vessel disease. vii, xii, xv, xxvi, 8, 11, 12, 33, 34, 37, 38, 47, 49, 121–124, 126–129, 133–135

SVM support vector machine. xii, xxi, 15–17, 30, 31, 121, 123

TP true positive. xviii, 83, 84, 86, 90, 92, 93, 96

WM white matter. 35, 46

WML white matter lesion. xvii, xxii, xxiii, 5, 8, 9, 11, 33–49, 131

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Medical Imaging Overview

Medical imaging techniques can generate detailed images representing the human anatomy in vivo [16]. The generated images reveal structural and functional information about organs and tissue. This information can be used to assist clinical diagnosis and interventions. The process of medical imaging is typically noninvasive. This means no instrument cuts the skin and is inserted into the patient's body. However, in some cases contrast agents are administered to reveal structural or functional information. The most common medical imaging modalities of interest for this thesis include radiography, CT, MRI and ultrasonography (ultrasound).

1.1.1 Imaging Modalities

Radiography: Radiography uses X-rays to visualize the human anatomy [17]. Specifically, a generator produces beams of X-rays and projects them to human bodies. Different organs and tissue absorb different amount of energy of the X-rays depending on their densities and compositions. The remaining X-ray energy is then captured by a detector and used to create an 2D image.

There are two types of radiographic images, namely projection radiography and fluoroscopy. Projection radiographs are also known as X-rays. They are widely used to diagnose and assess mus-

culoskeletal diseases as well as lung diseases. In fluoroscopy, dynamic projection radiographs are required. This can be used to visualize vessels via contrast agents that have been injected, as well as instruments such as catheters which are used to guide interventions.

Even though radiography is the oldest medical imaging technique developed, it is widely used because of its wide availability and low cost. However, radiography exposes patients and staff to harmful X-ray radiation.

CT: X-ray CT is based on the same principle as the projection radiography. However, in CT, a series of projectional X-ray images is acquired while the X-ray source and detector rotate around the patient [18]. The acquired data can be used to measure the X-ray attenuation coefficient at every point within a cross-section through the patient's anatomy. Subsequently, the coefficient is linearly transformed to the Hounsfield unit (HU). The HU provides a quantitative scale for radiodensity description. For a CT image, HU values are defined as:

$$HU = \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \times 1000. \tag{1.1}$$

Here, μ is the average linear attenuation coefficient of the voxel; μ_{water} and μ_{air} are radiodensities of distilled water and air at standard pressure and temperature, respectively. μ_{water} is defined as 0 and μ_{air} is defined as -1000. As a result, the intensities of voxels in a CT image typically range from -1000 to 1000.

CT is used in diagnosis in pathologies of a number of organs, including brain, heart, and lung. This is because CT images provide high resolution as well as good soft tissue contrast. Figure 1.1(a) shows a brain CT image.

MRI: MRI uses magnetic fields, gradients, and radio waves to create tomographic images of the patient's anatomy [19]. Briefly, pulses of radio waves excite the nuclear spin energy transition in hydrogen atoms and this is spatially localized by magnetic field gradients. Changing the parameters of the pulse sequence can generate different contrast mechanisms between tissues. The most commonly used imaging sequences include: T1-weighted, T2-weighted, DWI, dynamic contrast enhancement, and spectroscopy.

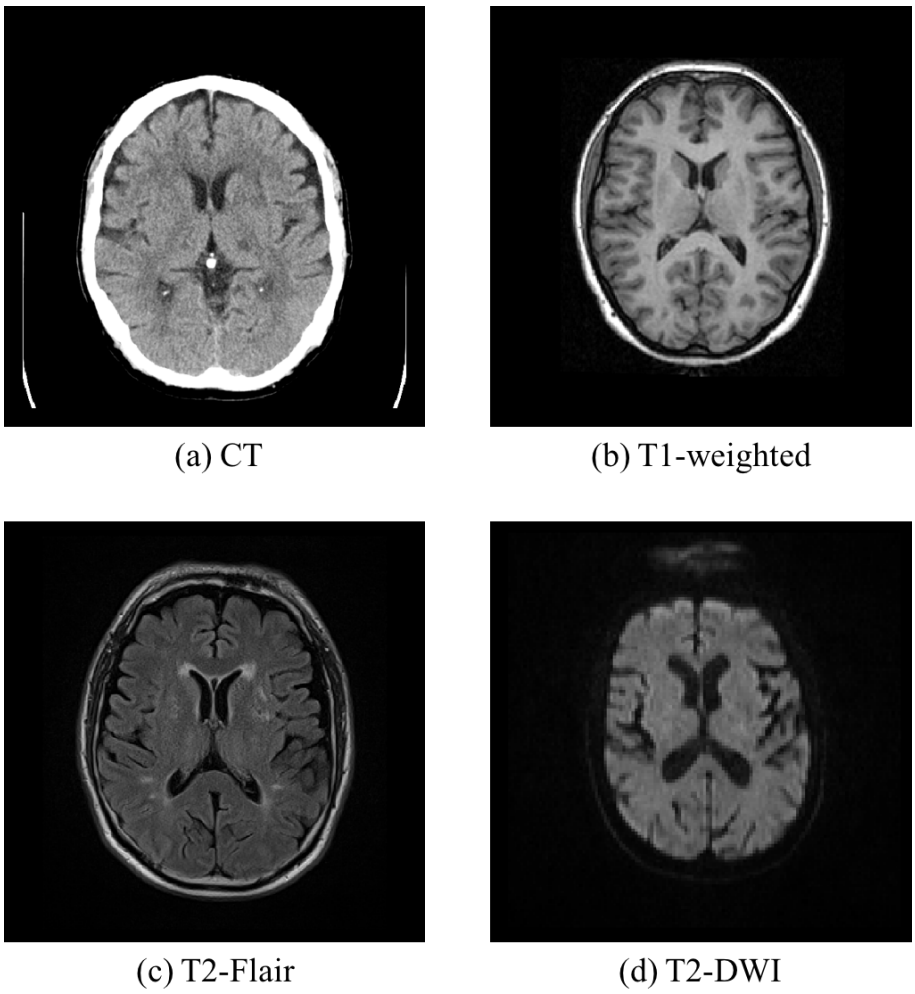


Figure 1.1: Examples of brain CT and MR images. The MR images include T1-weighted, T2-FLAIR, and T2-DWI sequences.

Compared with CT, MRI does not use X-rays or ionizing radiation, which is not harmful to human bodies. However, the acquisition of MRI is usually slower than CT imaging. MRI cannot be used in patients with non-removable metal implants as this causes safety issues. In addition, MRI offers better and more detailed soft tissue contrast than CT while the contrast between soft tissue and bone is higher in CT images. Figure 1.1 compares brain images acquired using different modalities. It is obvious that MR scans show better contrast between different brain tissues compared to the CT scan.

Ultrasound: Ultrasonography uses high frequency sound waves to generate 2D or 3D images [20]. Pulses of ultrasound waves are generated and sent into the patient's body by the ultrasound probe. The sound waves are reflected at the interface between different tissues. The ultrasound probe then records the reflected soundwaves in order to reconstruct images from them.

Ultrasound is widely used in imaging fetus in pregnant women, abdominal organs and the heart, because there is no harmful radiation and the cost is low. Furthermore, ultrasound images can be acquired in real-time. However, ultrasound images are usually of poor quality because of noise, artefacts, and shadows. Figure 1.2 shows an ultrasound image of a fetal brain.

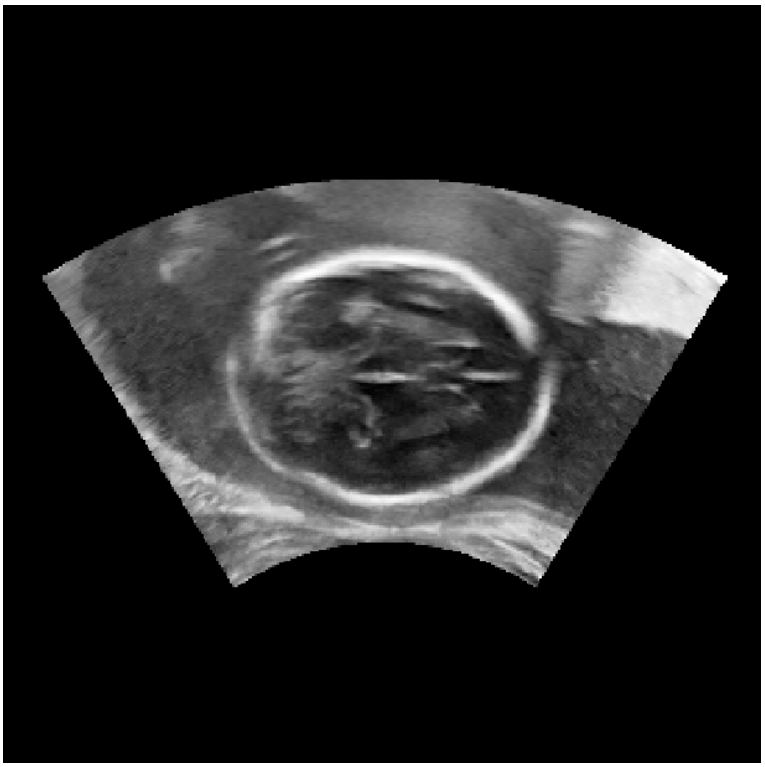


Figure 1.2: An example of ultrasound fetal brain image.

Others: There are many other medical imaging modalities which are used in diagnosis and interventions: positron emission tomography (PET), single-photon emission computed tomography (SPECT), elastography, endoscopy, tactile imaging, thermography as well as optical imaging. As this thesis does not use these techniques, they are not discussed in detail here. For more details on these medical imaging techniques, see [21].

1.1.2 Image Analysis and Interpretation

In clinical practice, the acquired images are subsequently interpreted by radiologists and clinicians. According to the features they identify in the images, different diseases can be diagnosed and treatments can be planned. The features of interest can be divided into two types, namely background

features and disease-specific features. For instance, when reading a brain CT scan of a stroke patient, a radiologist pays attention to background features, including white matter lesions (WMLs), atrophy as well as old lesions. The radiologist then focuses on disease-specific features (in this case stroke) such as acute ischemic lesions. These features are usually interpreted as scores, which indicate the severity of the disease. For instance, the Wahlund [22] and van Swieten [23] proposed scoring systems which include four and three grades of WML severity, respectively.

The process of clinical image analysis can be separated into three steps: First, identify regions of interest, e.g. identifying CT image slices with lesions. Second, delineate the target structures of interest, e.g. lesions. Finally, quantitative or qualitative measurements can be derived from the target structures of interest (e.g. size, shape or texture).

Analysing medical images automatically has three major advantages: First, compared to manual image analysis, automated algorithms can run faster, particularly those based on deep learning technologies, e.g. CNNs. For instance, the segmentation of acute ischemic lesion as developed in this thesis takes only seconds when performed automatically [24]. Second, inter- and intra-rater consistency of human experts can be low in challenging segmentation problems, e.g. WML segmentation in CT images [25]. Especially in the case of images with low signal-to-noise ratio or images with artefacts, automated methods can be derived from annotations provided by a committee of experts. Therefore, the results given by automated methods can be more reliable than those obtained from individual experts. Finally, it is expensive in terms of man power to annotate medical images by experts while automated analysis is more scalable.

Machine learning techniques have been used in automatic medical image analysis for decades. Traditionally there were only limited computational resources. In addition, there have been a small number of images available. Hand-crafted features such as the scale-invariant feature transform keypoints [26], were used as the input to machine learning models. The output of the models were targeted outcomes (e.g. labels, images) in supervised and weakly supervised learning or data intrinsic structures in unsupervised learning. In recent years, many more images and annotations as well as computational resources have become available. Image intensities can be input to machine learning models, particularly deep neural networks. The deep neural networks can learn image features and

subsequent tasks end to end, which improves the performance significantly.

1.2 Objective and Challenges

The main objective of this thesis is to develop machine-learning-based models to solve the classification and segmentation problems assisting stroke clinics. The developed methods can also be applied to other medical imaging problems.

1.2.1 Medical Images

The nature of medical images leads to three major challenges in medical image analysis. First, in some cases, the acquired images have poor quality. Specifically, clinical images are usually optimized for diagnostic purposes, e.g. to minimize the radiation burden for patients or maximize the acquisition speed. This can result in images with low signal-to-noise ratio. In addition, clinical images are often degraded by artefacts, e.g. due to patient movement or natural motion (e.g. respiratory or cardiac motion).

Second, the diversity of medical images is enormous. More precisely, there are many different image modalities mentioned above which lead to very different images in terms of appearance and intensity distributions. Even for a single modality, such as CT, images of different organs are likely to have very different intensity distributions. However, common problems happen in different medical images. For instance, segmentation is a common problem in medical image analysis, e.g. brain segmentation in CT and MR images. Developing individual models for each case based on machine learning is problematic and redundant. It is significant to develop generic models for common problems.

Finally, the human anatomy and its appearance in medical images can be complex and highly variable in terms of structure, position, size, and shape. In terms of structure, some anatomies can be seen as a cluster of pixels (or voxels), i.e. blobs, such as tumours [27], while some of them look like a mass of pixel (or voxel) dots, such as micro-bleeds [28]. In terms of position, lesions, e.g. ischemic infarcts, can occur everywhere in the brain [24]. In terms of size, some organs are small, e.g. pancreas, while

some organs are large, e.g. liver. In terms of shape, it depends on anatomy itself and the pose it is scanned.

1.2.2 Annotations

Annotations on medical images can exist across three levels: image level, object level, and pixel (or voxel) level. From image level to pixel (or voxel level), the annotation becomes more and more expensive in terms of human resource and time. Because of this, two challenges raise in terms of image annotations.

First, only a small portion of images can be annotated. Acquiring a large number of clinical scans from hospitals is not difficult but annotating them at pixel (or voxel) level is often not possible due to resource constraints. For segmentation problems, it is common that experts only annotate a small number of images for model development and validation. This means that many images remained unlabelled. It is challenging to make use of these images without annotations to boost the model performance.

Second, the images can often be weakly annotated. This means only course-grained annotations are provided to solve fine-grained problems. For instance, only image-level annotations are given for detection or segmentation problems. In this case, there are two possible solutions: One is to derive fine-grained annotations from course-grained annotations; the other is to transform the fine-grained problem to a course-grained problem. Both of them pose significant challenges.

1.2.3 Clinical Applications

Stroke is a cerebrovascular accident which is the loss of brain function caused by the lack of blood supply [29]. It is one of the major causes of long-term disability and death globally [30]. Ischemic stroke and hemorrhagic stroke are two different categories of strokes that require different treatments [31]. Ischemic stroke accounts for approximately 80% of all strokes [32]. A number of factors such as energy depletion and cell death are thought to result in ischemic brain injuries [33]. Intravenous

thrombolysis with recombinant tissue plasminogen activator is the recommended therapy for acute ischemic stroke that reduces severe disability but causes deterioration due to symptomatic intracranial hemorrhagic (ICH) in approximately 6% [34]. In order to reduce the rate of ICH which is associated with the worst outcome of stroke, management of ischemic stroke is pivotal.

Advanced neuroimaging techniques have been widely used in the diagnosis of stroke. It is normally recommended that patients should undergo either MRI or CT [35]. DWI and T2-FLAIR should be included in the MR sequences which are able to show acute and chronic lesions, respectively. Although MRI is regarded as the gold standard, CT is more frequently used in the acute phase of stroke treatment since CT is more widely applicable and faster.

In this thesis, a number of key biomarkers associated with ischemic stroke are identified and analyzed, including the SVD, atrophy, and acute ischemic lesions. In stroke clinics, the SVD and atrophy are recognized as background biomarkers while the acute ischemic lesions are the acute biomarkers [34, 36]. Cerebral SVD refers to a group of pathological aetiologies that affect the brain [37]. In this thesis we will use this term to describe ischemic consequences of WMLs. The CSF volume is a biomarker of atrophy. Quantifying CSF volume within ventricles and cortical sulci is significant for distinguishing hydrocephalus from central atrophy [38], for prognostication after stroke [39], and for estimating cerebral hemorrhage risk [40, 41]. In addition, accurately detecting the acute ischemic lesions in medical images directly contributes to the stroke diagnosis. For instance, small ischemic lesions which are likely to be missed by clinical observers can be highlighted. The efficiency of scan reviewing can be boosted as well.

Therefore, the methods developed in this thesis are applied in the following stroke-related problems: 1) rapid identification of cerebral SVD on CT images (Chapter 7), 2) accurate segmentation of cerebral SVD on CT images (Chapter 3), 3) multi-class segmentation of CSF on CT images (Chapter 4), and 4) acute ischemic lesion segmentation on DW images (Chapter 5).

1.3 Contributions

In this thesis, we propose novel solutions based on machine learning techniques to address the challenges mentioned above.

A fully automated framework is developed for the segmentation of the WMLs in brain CT images with often poor quality. Assessment of cerebral ischemic WMLs (or leukoaraiosis) using CT is important for the practical management of acute stroke, traumatic head injury and cognitive impairment, but limited by visual rating systems that are often used but prone to ambiguity and high inter-rater variability. We propose a framework based on the random forests algorithm to segment the WMLs so that the lesions can be quantified reliably. Image patches across multiple scales are used to address the challenge that CT images can exhibit poor quality. We demonstrate that the automatically calculated WML volumes strongly correlate to WML volumes derived from expert drawings on MRI and CT ($r^2 = 0.85, 0.71$, respectively; $p < 0.001$). Expert CT-WML drawing volumes correlated with each other ($r^2 = 0.85$), but ranged widely between experts (range: 91% of mean expert estimate). Agreements between automatic and consensus-expert score ratings were superior or similar to agreements between pairs of experts. Accuracy was unaffected by co-existent old or acute ischemic changes, or atrophy. Automatic rating errors (scores > 1 point from expert consensus) occurred in 4% cases.

A generic CNN architecture is proposed for segmentation problems in medical imaging. The U-Net architecture [6] is one of the most well-known CNN architectures for semantic segmentation and has achieved remarkable successes in many different medical image segmentation applications. It consists of standard convolution layers, pooling layers, and upsampling layers. These convolution layers learn representative features of input images and construct segmentations based on the features. However, the features learned by standard convolution layers are not distinctive when the differences among different categories are subtle in terms of intensity, location, shape, and size. We propose a novel CNN architecture, called Dense-Res-Inception Net (DRINet) with deeper and wider layers, which addresses this challenging problem. The proposed DRINet consists of three blocks, namely a convolutional block with dense connections, a deconvolutional block with residual Inception modules, and an unpooling block. Our proposed architecture outperforms the U-Net architecture [6] in

three different challenging applications, namely multi-class segmentation of CSF on brain CT images, multi-organ segmentation on abdominal CT images, multi-class brain tumour segmentation on MR images.

A CNN-based framework is developed for the segmentation of the complex ischemic lesions in brain DW images. Stroke is an acute cerebral vascular disease, which is likely to cause long-term disabilities and death. Acute ischemic lesions occur in most stroke patients. These lesions are treatable using drugs provided that an accurate diagnosis is available. Although DWI is sensitive to these lesions, localizing and quantifying them manually is costly and challenging for clinicians since the lesions significantly vary in location, size, and shape. We propose a novel framework to automatically segment stroke lesions in DWI. Our framework consists of two CNNs: one is an ensemble of two DeconvNets [2], which we term EDD Net; the second CNN is the multi-scale convolutional label evaluation net (MUSCLE Net), which aims to evaluate the lesions detected by the EDD Net in order to remove potential false positives. Our proposed framework is validated on a large dataset comprising clinical acquired images from 741 subjects. A mean accuracy of Dice coefficient obtained is 0.67 in total. The mean Dice scores based on subjects with only small and large lesions are 0.61 and 0.83, respectively. The lesion detection rate (DR) achieved is 0.94.

A self-supervised learning strategy is proposed for CNN pretraining, which improves the performance of CNN. Machine learning, particularly deep learning has boosted medical image analysis over the past years. Training a good model based on deep learning requires large amount of labelled data. However, as mentioned above it is often difficult to obtain a sufficient number of labelled images for training. In many scenarios, the dataset in question consists of more unlabelled images than labelled ones. Therefore, boosting the performance of machine learning models by using unlabelled as well as labelled data is an important but challenging problem. Self-supervised learning presents one possible solution to this problem. However, existing self-supervised learning strategies applicable to medical images do not result in significant performance improvement. In this thesis, we propose a novel self-supervised learning strategy based on context restoration, i.e. restoring randomly disordered image context, in order to better exploit unlabelled images. The context restoration strategy has three major features: 1) it learns meaningful image semantics; 2) it is useful for different types of subsequent image analysis tasks; and 3) its implementation is simple. We validate the context

restoration strategy in three common problems in medical imaging: classification, localization, and segmentation. For classification, we apply and test it to scan plane detection in fetal 2D ultrasound images; to localise abdominal organs in CT images; and to segment brain tumours in multi-modal MR images. In all three cases, the proposed self-supervised learning based on context restoration learns meaningful semantic features and leads to improved machine learning models for the above tasks.

A multi-instance learning (MIL)-based method is proposed for the identification of cerebral SVD with weak labels. Cerebral SVD is a common cause of ageing-associated physical and cognitive impairment. Identifying SVD is important for both clinical and research purposes but is usually dependent on radiologists' evaluation on brain scans. CT is the most widely used brain imaging technique but for SVD it usually has a low signal-to-noise ratio, and consequently low inter-rater reliability. The SVD is only related to regions affected by the disease but these regions are not annotated. The annotations are based on image level, i.e. absent/mild SVD or moderate/severe SVD. We propose a novel framework based on MIL to distinguish between absent/mild SVD and moderate/severe SVD. Intensity patches are extracted from regions with high probability of containing lesions using an atlas-based approach. These are then used as instances in MIL for the identification of SVD. A large baseline CT dataset, consisting of 590 CT scans, was used for evaluation. We achieved approximately 75% accuracy in classifying two different types of SVD, which is high for this challenging problem. Our results outperform those obtained by either standard machine learning methods or current clinical practice.

1.4 Structure of Thesis

The remainder of this thesis is structured as follows: Chapter 2 introduces those machine learning techniques that relevant to the work in this thesis. In Chapter 3, we propose a framework based on RF to segment and quantify the WMLs in clinical CT images. In Chapter 4, a generic CNN architecture is proposed for medical image segmentation problems, which is shown to be robust across image modalities and different segmentation problems. In Chapter 5, we propose a CNN-based framework for acute ischemic lesion segmentation. In this application, the lesions vary in position, size, and

shape, making this a challenging problem. In Chapter 6, a novel self-supervised learning strategy is proposed. The proposed strategy improves the performance of CNNs where only limited annotated images are available. In Chapter 7, a framework based on MIL is proposed to distinguish different types of SVD in CT images. In this case, the SVD is only related to regions affected by the disease but the regions are not annotated. Finally, we conclude our work and discuss the limitations and future plans in Chapter 8.

Chapter 2

Background

2.1 Introduction

Machine learning (ML) is a subject which uses empirical data $\mathbf{X}$ and computational models $\mathcal{M}$ to approximate a function, e.g. $f(\cdot)$. The empirical data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ consists of m instances and each data instance $\mathbf{x}_i = \{x_1, x_2, \ldots, x_n\}, i = 1, 2, \ldots, m$, can be viewed as a vector of n feature attributes. The function to be approximated is usually highly complex and implicit. According to whether the training data has associated labels $\mathbf{Y}$ available, ML techniques can be categorized into two categories, namely supervised learning and unsupervised learning. In supervised learning, the training data has associated labels $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ while in unsupervised learning the data is unlabelled, i.e. $\mathbf{Y} = \mathbf{X}$. If only a portion of data (m_p/m instances) is annotated, i.e. $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{m_p}\}$,, the learning process is called weakly supervised learning. Commonly the learning process aims to find an approximation of function $f : \mathbf{X} \mapsto \mathbf{Y}$.

The major problems in supervised learning are classification and regression. In the classification problems, $\mathbf{y}_i, i = 1, 2, \ldots, m$, is discrete and categorical while in regression problems, $\mathbf{y}_i, i = 1, 2, \ldots, m$, is usually continuous. The main application for unsupervised learning are clustering and dimensionality reduction. In clustering problems, the result of the learning process is a set of clusters while in dimensionality reduction problems, the result of the learning process is a representation of original data instances in a lower dimensional space. In weakly supervised learning, the aim is to make use of

data without labels to improve the performance of supervised learning tasks. In reinforcement learning, the goal is to achieve maximum rewards via an optimized policy. The challenge is that there is no "instance-label" pairs for policy learning. A reward is only available after a sequence of actions.

In medical image analysis, the most common and challenging problem is classification, e.g. classifying subjects into disease categories or classifying image pixels (or voxels) into different classes according to tissue types or organs. In cases where all images are labelled (or annotated), supervised learning algorithms are applicable. However, in classification problems, there are many images which are not labelled because of prohibitive costs of annotating images. Therefore, weakly supervised methods are utilized to make use of these unannotated images.

2.2 Supervised Learning

2.2.1 Basic Models

In terms of supervised learning algorithms, linear models such as the logistic regression and probabilistic models such as the naive Bayes classifier are basic building blocks often used in classification tasks [42]. A linear model

$$z = \mathbf{w}^T \mathbf{x}_i + b, \tag{2.1}$$

predicts z as a linear function of $\mathbf{x}_i$ Here, $\mathbf{w}$ and b are learned parameters of the linear model. The predicted variable z could be converted to a categorical number y_i using the unit step function

$$y_i = \begin{cases} 0, & z < 0; \\ 1, & z \geq 0. \end{cases} \tag{2.2}$$

However, the unit step function is not continuous. Therefore, the logistic function:

$$y_i = \frac{1}{1 + e^{-z}} \tag{2.3}$$

is usually used to replace the unit step function. Instead of predicting the classes directly, the logistic function aims to predict the odds of each class, which can be more useful. This linear model is a discriminative model, which estimates the posterior probability of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $P(\mathbf{Y} \mid \mathbf{X})$ directly. This can also be achieved via generative models. Generative models estimate the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $P(\mathbf{X}, \mathbf{Y})$ and

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{P(\mathbf{X}, \mathbf{Y})}{P(\mathbf{X})}. \tag{2.4}$$

According to Bayes' theorem,

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{P(\mathbf{Y})P(\mathbf{X} \mid \mathbf{Y})}{P(\mathbf{X})}. \tag{2.5}$$

Here, $P(\mathbf{Y})$ is the prior probability of $\mathbf{Y}$; $P(\mathbf{X} \mid \mathbf{Y})$ is the likelihood of $\mathbf{X}$ belonging to $\mathbf{Y}$ class; and $P(\mathbf{X})$ is the normalization evidence. However, it is difficult to estimate the likelihood $P(\mathbf{X} \mid \mathbf{Y})$ via limited training instances. The naive Bayes classifier assumes all attributes are conditionally independent given $\mathbf{Y}$. Therefore,

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{P(\mathbf{Y})}{P(\mathbf{X})} \prod_{i=1}^{M} P(\mathbf{x}_i \mid \mathbf{y}_i). \tag{2.6}$$

2.2.2 SVM

Another approach for classification is based on support vector machines (SVM) [43]. The standard SVM solves the binary classification problem, where $\mathbf{y}_i \in \{-1, +1\}, i = 1, 2, \ldots, m$. It aims to find a hyper-plane

$$\mathbf{w}^T\mathbf{x} + b = 0, \tag{2.7}$$

which separates the data instances into two classes. Here, $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$ is the normal vector and b is the bias. The optimal hyper-plane should be at the centre of the maximum margin between instances of difference classes. Figure 2.1 demonstrates the maximum margin in a 2D example.

Therefore, the problem to solve can be written as:

$$\max_{\mathbf{w},b} \quad \frac{2}{\|\mathbf{w}\|} \tag{2.8}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, m. \tag{2.9}$$

This optimization problem equals to

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{2.10}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, m. \tag{2.11}$$

This can be solved using the method of Lagrange multipliers.

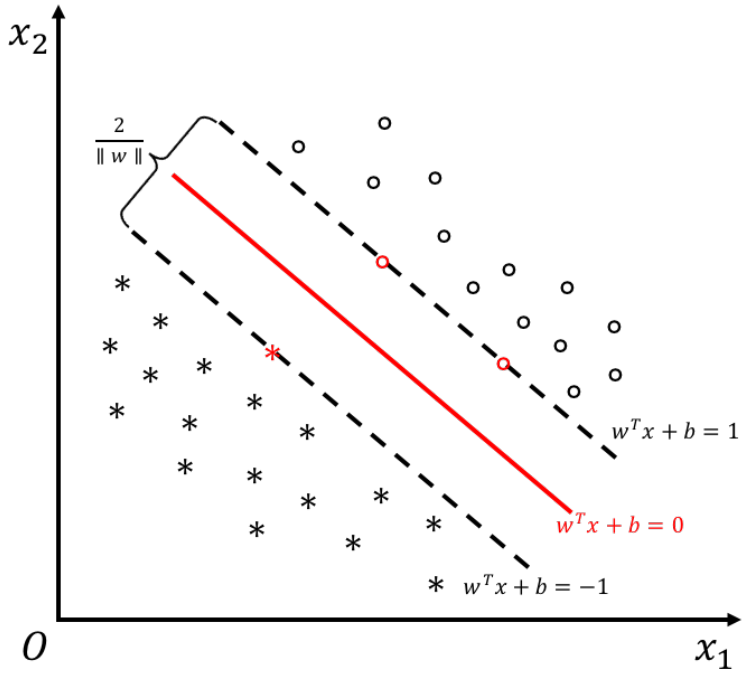


Figure 2.1: A demonstration of margins and support vectors in the SVM approach. Here, the circles and stars represent positive and negative instances belonging to different classes, respectively. The instances in red are support vectors.

Note that the model of the hyper-plane in the SVM is linear:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b. \tag{2.12}$$

However, these models cannot solve problems which are not linearly separable (e.g. the XOR prob-

lem). If there is a function $\phi(\cdot)$, which can map the data into a latent space, where the results are linearly separable, the SVM model still can be used. In this case, the model of the hyper-plane is:

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b. \tag{2.13}$$

This is known as the kernel trick which maps a linear classifier to a non-linear classifier via a kernel function $\phi(\cdot)$.

To alleviate the effect of overfitting, soft margins are usually applied, which means some errors are allowed. In this case, the constraints in Equation 2.11 can be relaxed and the optimization problem becomes:

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{2.14}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, m, \xi_i \geq 0. \tag{2.15}$$

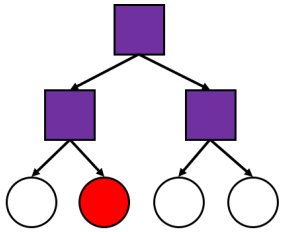
Here, C is a constant greater than 0 and ξ_i are referred to as slack variables, which quantify the degree that data instances do not satisfy the constraints 2.11.

2.2.3 Random Forests

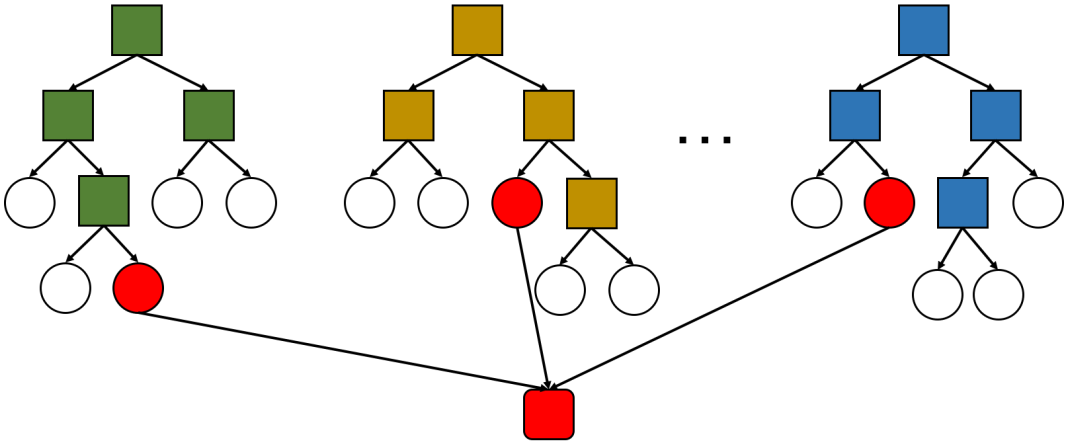
Random forest is an ensemble method based on decision trees [44]. Decision trees are a type of popular machine learning method which are regarded as weak learners in RF. Figure 2.2 shows examples of a decision tree and a random forest. Generally, a decision tree consists of one root node and a number of internal nodes and leaf nodes. Each leaf node corresponds to a decision while all the other nodes correspond to feature thresholding. Each non-leaf node contains a subset of data samples. The data samples are then divided to its child nodes according to the feature thresholding results. The root node contains the whole dataset. The key of developing a decision tree is the strategy to divide the feature space. The goal of feature dividing is to maximize the purity of each non-leaf node, which means as many data samples as possible belonging to one non-leaf node are from one category.

Ideally, different weak learners should be independent in order to achieve the best performance after

combination. In practice, it is impossible to ensure the independence among weak learners. The alternative is to make weak learners as diverse as possible. One strategy is sampling subsets of data when training the different weak learners. In addition, the RF introduces another strategy to create diversities among weak learners: Specifically, for each node in a decision tree, the optimal criterion for division is obtained from k randomly selected attributes among all feature attributes $x_i, i = 1, 2, \ldots, n$. If $k = n$, then the tree construction is equal to standard decision tree construction; while if $k = 1$, then the selection of attribute for division is completely random. This means k controls the randomness and usually $k = \log n$ or $k = \sqrt{n}$ [44]. There are several methods for combining weak decision trees in form of RF, including averaging [45], majority voting, and learning-based methods [46–48].



(a) An example of decision tree



(b) An example of random forest

Figure 2.2: Comparison between a decision tree and a random forest. A random forest consist of a number of decision trees. The output of the random forest (the rounded rectangle in red) is a combination of the outputs of its decision trees. In each decision tree, rectangles and circles represent internal nodes and leaf nodes, respectively. Rectangles in different colours suggest different attribute sets. Circles in red indicate the outputs of decision tress.

2.2.4 Neural Networks

The Origin

In 1943, McCulloch and Pitts (M-P) proposed a type of neural networks, the M-P neuron model [49], shown in Figure 2.3. In the M-P neuron model, n input signals $x_j, j = 1, 2, \ldots, n$ are summed with associated weights w_j. The weighted sum is then compared with a threshold θ. The result is subsequently processed by a activation function $g(\cdot)$, leading to the output $y_i, i = 1, 2, \ldots, m$. Formally,

$$y_i = g\left(\sum_{j=1}^{n} w_j x_j - \theta\right). \tag{2.16}$$

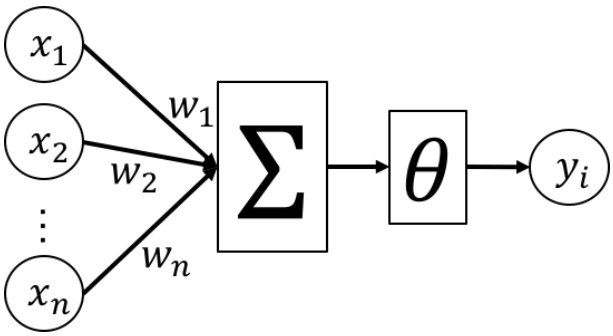


Figure 2.3: M-P neuron model

The activation function is the step function:

$$g(x) = \begin{cases} 0, & x < 0; \\ 1, & x \geq 0. \end{cases} \tag{2.17}$$

However, the step function is not continuous so the sigmoid function [50] is usually used instead. A sigmoid function has a characteristic "S"-shaped curve, which is bounded and differentiable.

In 1958, Rosenblatt invented the perceptron [51], which consists of two layers of neurons, namely the input layer and the output layer. The input layer receives input signals and the output layer is a M-P neuron model. In the perceptron, the weights w_j and the threshold θ can be learned. Since the threshold can be viewed as a fixed value $x_{n+1} = 1$ with weight w_{n+1}, the target is to learn all the weights of the model. The rule of learning is fairly simple: Let the output of perceptron is $\bar{y}_i$ given

the input $(\mathbf{x}_i, y_i)$. If $\bar{y}_i = y_i$, then the weights do not change; otherwise,

$$w_j \leftarrow w_j + \alpha(y_i - \bar{y}_i)x_j. \tag{2.18}$$

Here, $\alpha \in (0, 1)$ is the learning rate. The perceptron is able to handle linearly separable problems because it only has one layer of functional neuron. To solve the XOR problem and non-linear problems, more layers of functional neurons are needed. However, in this case Equation 2.18 is no longer applicable. The backpropagation (BP) algorithm is the most successful algorithm to solve the multi-layer network learning problem. The BP algorithm was first proposed by Werbos [52] and popularized by Rumelhart et al. [53]. The BP algorithm computes the weights based on the chain rule to find the gradients of the loss function.

Modern Neural Networks

In theory, neural networks with large number of neurons can solve highly complex problems. Particularly, increasing hidden layers leads to better results than increasing neurons in existing hidden layers. This is because more hidden layers lead to deeper embeddings of non-linear activations. In practice, training complex neural networks is difficult due to the problem of vanishing gradients with increasing network depth [54]. One successful strategy to solve this problem is weight sharing, which means a groups of neurons share the same weights, leading to the development of CNN. In a CNN, one filter works on a pair of connected feature maps with a number of neurons [55].

In CNNs, convolutions and poolings are two key components of the neural network. The idea of convolutions and poolings was inspired by a study on the visual system of cats [56]. Fukushima and Miyake first proposed to use convolutions and poolings in neural networks [57, 58]. Modern CNNs were proposed in [55] which applied the BP algorithm on neural networks. A CNN usually consists of stacks of convolutional layers and pooling layers. In each convolution layer, there are a number of feature maps, which are groups of neurons. The features are extracted by convolution filters. The feature maps are downsampled in pooling layers, which are based on local features in feature maps. The pooling layers remove redundancies in feature maps, which improves the learning efficiency.

The convolution and pooling layers extract features from the input layer-by-layer, resulting in more and more representative features. Ultimately, these features can be used for classification via simple classifiers.

In 1998, LeCun et al. proposed the LeNet-5, which achieved a great success in the hand-written document recognition task [7]. Figure 2.4 shows the architecture of the LeNet-5. Details about the implementations including the weight updating in the CNN can be found in [59]. In the following years, a number of techniques were proposed to improve the CNN architectures. In terms of the activation function, the ReLU

$$g(x) = \max(0, x) \tag{2.19}$$

was proposed to replace the sigmoid function to avoid gradient vanishing and accelerate the gradient computation [60]. To prevent neural networks from overfitting, a technique called dropout was proposed [61,62]. The dropout technique sets the outputs of hidden neurons to 0 with a certain probability, e.g. 0.5. The dropped out neurons are not involved in the forward inference pass or during BP. This means the whole network samples different architectures for training and all these architectures share weights. This strategy prevents overfitting since the learned weights need to adapt different structures, which forces the network to be robust.

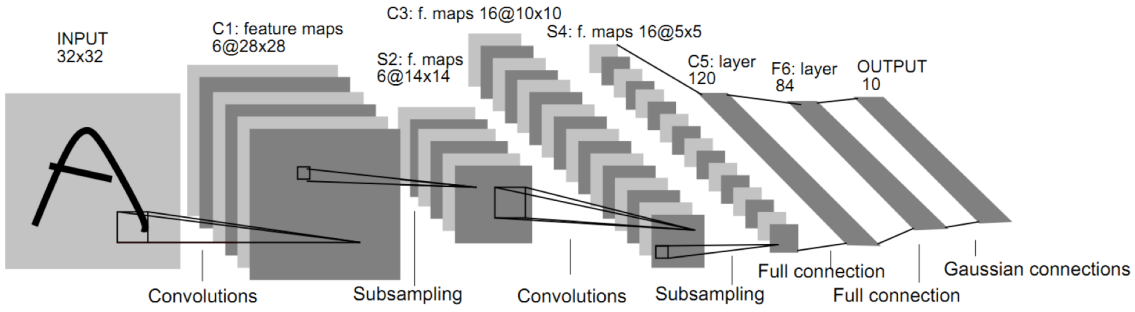


Figure 2.4: The architecture of the LeNet-5. The figure is from [7].

Applications

Using these techniques, Krizhevsky et al. proposed the AlexNet [8], shown in Figure 2.5, which won the ImageNet large scale visual recognition challenge (LSVRC)-2010 contest [63] with a significantly increased accuracy (top-5 error rate of 15.3%). The AlexNet has five convolution layers and three

fully-connected layers. Later, the VGGNet was proposed [64], which has two similar architectures with 16 layers and 19 layers, respectively. The configuration of these VGGNet architectures is shown in Table 2.1. The VGGNet achieved the top-5 error of 6.8% on the classfication problem of LSVRC-2014. This result suggested deeper networks lead to better results. The GoogLeNet [9] achieved similar accuracies to the VGGNet but its architecture is deeper and wider. The architecture of the GoogLeNet is shown in Figure 2.6. The increase of depth and width of network architecture does not add many more parameters because of the careful design of the inception modules proposed. In the inception module, convolutions with 1×1 kernels are used, which limits the number of feature maps (see details in Chapter 4).

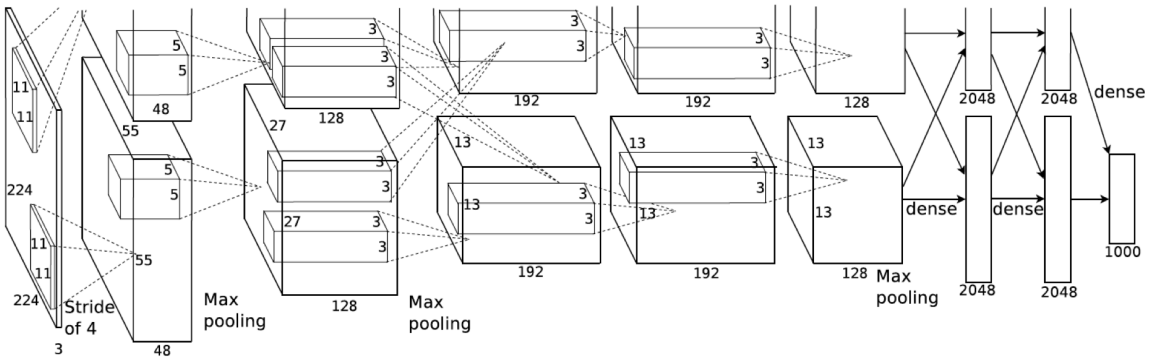


Figure 2.5: The architecture of the AlexNet. The figure is from [8].

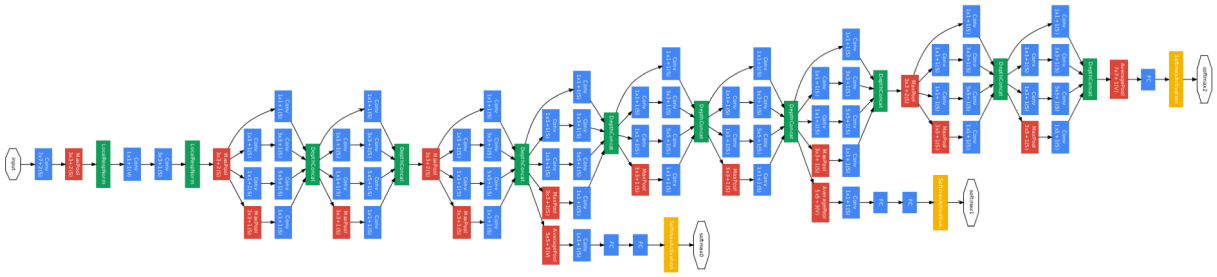


Figure 2.6: The architecture of the GoogLeNet. The blocks in blue, red, yellow, and green represent convolutions, poolings, softmax, and concatenations and normalization, respectively. The figure is from [9].

In 2016, a CNN with residual connections was proposed, termed ResNet [10]. The ResNet extends the VGGNet's depth to 34, 52, and 101 layers without introducing extra parameters. The architecture of the ResNet-34 is shown in Figure 2.7. The residual connections solve two training problems resulting

Table 2.1: The configuration of the VGGNets with 16 and 19 weight layers.

16 weight layers	16 weight layers	19 weight layers
input images		
conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
max pool		
conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
max pool		
conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-512
max pool		
conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
max pool		
conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
max pool		
fc-4096		
fc-4096		
fc-1000		
softmax		

in good performance: One problem is that of vanishing gradients, which means the errors in high layers are likely to vanish when backpropagating to low layers (layers close to the input are usually referred to as low layers while layers close to the output are referred to as high layers). The other problem is the degradation of training accuracy which is not caused by overfitting. The degradation is due to the difficulty in function approximation. If the optimal function is more likely to be an identity mapping than a zero mapping, it is easier to find the pertubations with reference to an identity mapping than to learn the function from scratch. The residual connections make the input feature maps as the reference. Hence, the learning process becomes easier. Similarly, there are convolutions with 1×1 filters within the residual connection blocks, which control the parameter space. Later, the DenseNet architecture [11] was proposed to improve the performance of the ResNet, where all preceding layers are connected to the following layers via concatenation to avoid vanishing gradients.

In the DenseNet architecture (Figure 2.8), the size of the output channel of a convolution layer is typically small (e.g. 12, 24). This is also referred to as the growth rate of the network which controls the parameter space. The dense connection pattern is demonstrated in Figure 2.9.

More recently, Sabour et al. proposed the CapsuleNet [12] to address the shortcomings of CNNs. This architecture is based on the observation that CNNs are not a good representation of the human visual system. The CNNs are translation invariant and require big data to generalize. The CapsuleNet learns a global linear manifold between a whole object and its pose in an unsupervised learning manner. In addition, in the CapsuleNet architecture, routing is dynamic, instead of using pooling layers. The dynamic routing means feature maps are forwarded to capsules which are the best at processing them. As such, the CapsuleNet is able to generalize better with less training data. A simple 3-layer CapsuleNet is demonstrated in Figure 2.10.

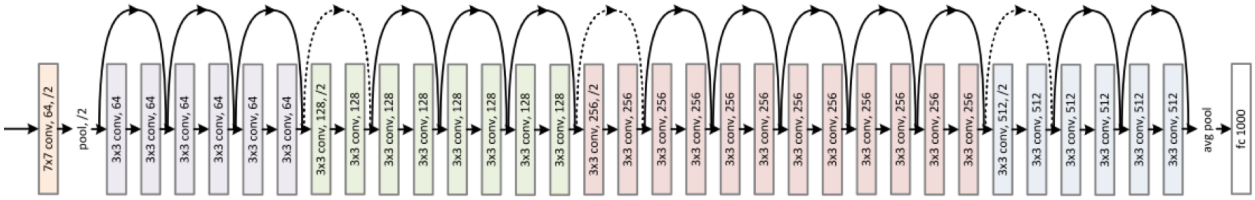


Figure 2.7: The architecture of the ResNet-34. The figure is derived from [10].

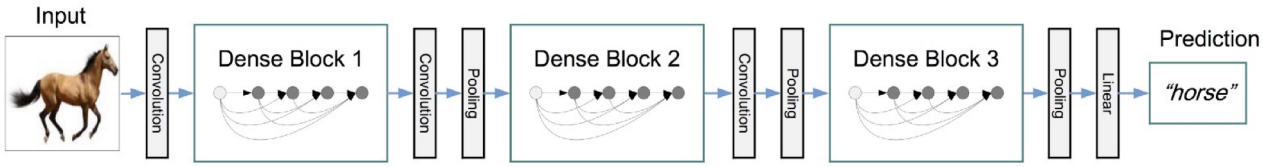


Figure 2.8: The architecture of the DenseNet. The figure is from [11].

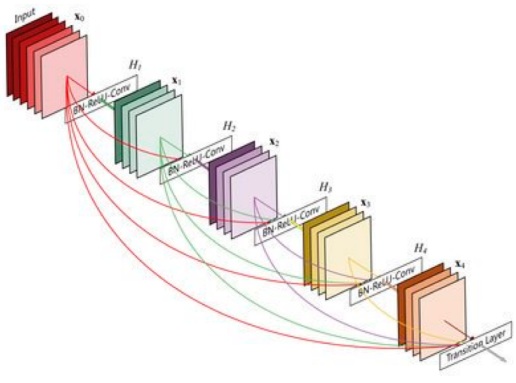


Figure 2.9: The demostration of the dense connection in the DenseNet. The figure is from [11].

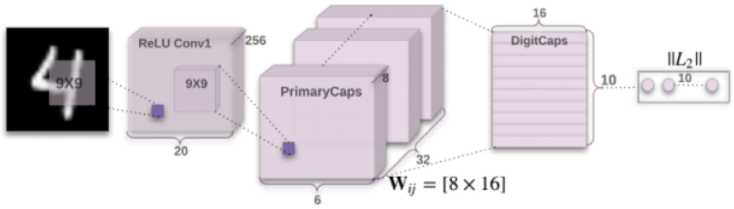


Figure 2.10: The architecture of a 3-layer CapsuleNet. The figure is from [12].

CNNs are often used to solve image classification problems, where a label is assigned to an image. In semantic segmentation problems, where a label is assigned to each pixel (or a voxel), CNNs can also be used. A CNN solving classification problems consists of two parts: The first part comprises convolution layers and pooling layers while the second part comprises fully-connected layers and a classifier. A modified CNN architecture suitable for semantic segmentation problems inherits the first part of the classification CNN while the second part usually consists of convolution layers, upsampling layers, and a classification layer, which generate a semantic segmentation map. The first part of the segmentation CNN is also referred to as the analysis path, extracting representative features from input images while the second part of it is referred to as the synthesis path, upsampling the feature maps from the analysis path and creating segmentation maps.

Long et al. [13] proposed the first CNN to address the segmentation problem, called FCN, shown in Figure 2.11. The FCN's analysis path is derived from the AlexNet, the VGGNet, and the GoogLeNet, respectively. The FCN's synthesis path combines feature maps from the analysis path across multiple scales. This is because high-level feature maps lose fine structures, which can be compensated for by using feature maps at lower levels.

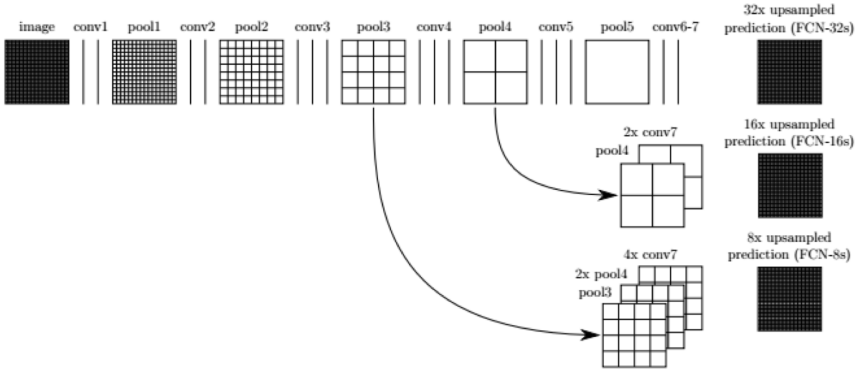


Figure 2.11: The architecture of the FCN proposed by Long et al. [13]. The figure is from [13].

The DeepLab architecture [14] is another commonly used CNN architecture for addressing segmenta-

tion problems. The VGGNet was employed as the backbone of its analysis path. However, to extract deeper features from input images without adding more parameters, convolutions with dilations were used to enlarge the field of view. In addition, atrous spatial pyramid pooling (ASPP) was proposed to aggregate feature maps achieved by convolutions with different atrous rates. As a result, the feature maps at the highest level of the DeepLab's analysis path assemble highly representative features of input images. These representative feature maps are then upsampled to the same size of the input image using bilinear interpolation, resulting in the raw segmentation map. Since the raw segmentation map is fairly coarse, a conditional random field (CRF) model is proposed to create the fine segmentation map. Figure 2.12 illustrates the DeepLab model. Note that the CRF refinement is not part of the end-to-end training process. Zheng et al. [65] formulated the iterations in CRF models as recurrent operations, which enables the end-to-end network training. This end-to-end training enables increased segmentation accuracy. Recently, the DeepLabV3 [66] was proposed. The atrous convolutions with multiple atrous rates were adopted, which encode objects in images at multiple scales. The ASPP was extended to include the image global features. These two major improvements boost the performance of the DeepLab model significantly.

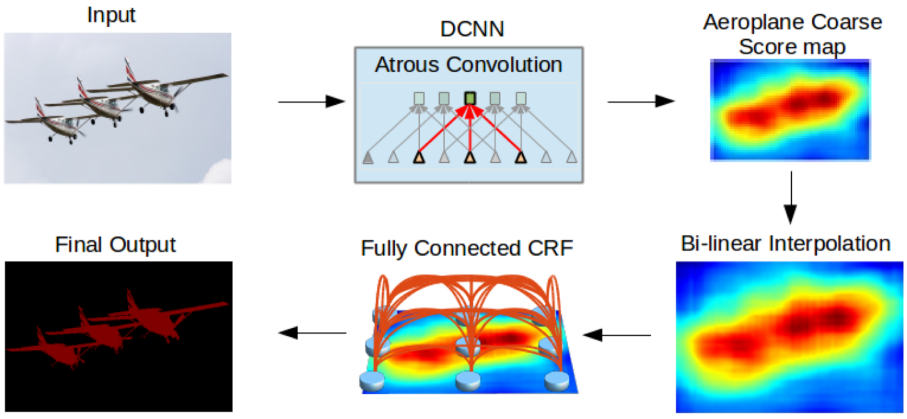


Figure 2.12: The illustration of the DeepLab model. Figure from [14].

The encoder-decoder architecture is another commonly used CNN architecture for semantic segmentation. This architecture is derived from the auto-encoder [3], where the analysis path and the synthesis path have symmetric convolution and deconvolution layers and pooling and upsampling layers. The SegNet [67] and the DeconvNet [2] are two representatives of this type of CNNs. Figure 2.13 shows the architecture of the DeconvNet. Unlike an auto-encoder, which reconstructs input images at the end, the SegNet and the DeconvNet create segmentation maps. Instead of using bilinear inter-

polation or deconvolution, the SegNet and the DeconvNet proposed a novel upsampling layer called unpooling layer. The unpooling layer records the locations of max activations in feature maps in max pooling layers and uses the recorded masks to guide the upsampling. This strategy improves the segmentation of object structures, which results in the improvement of total accuracies.

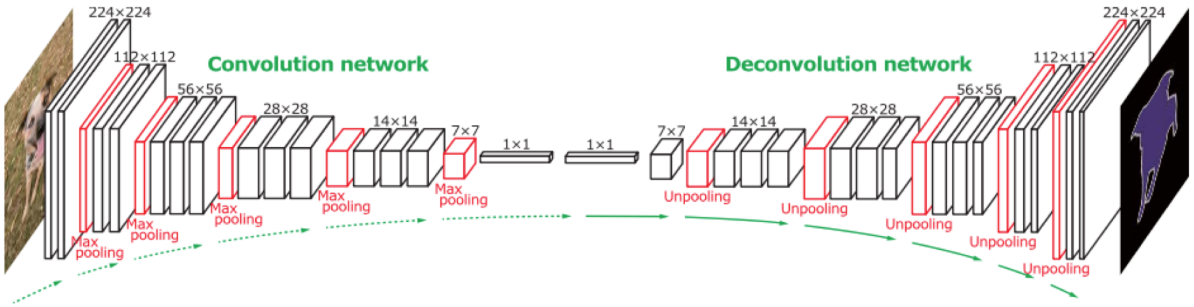


Figure 2.13: The architecture of the DeconvNet. The figure is from [2].

In addition to the encoder-decoder architecture, the U-Net architecture [6] proposes to connect the associated feature maps in the analysis path and the synthesis path via concatenation, which is illustrated in Figure 2.14. As such, gradients at high layers can be propagated to low layers directly, which alleviates gradient vanishing. Therefore, the overall performance of the network improves. In the conventional U-Net, the synthesis path comprises convolution layers and deconvolution layers. The deconvolution layers with a stride of 2 are used to upsample the feature maps. Beyond the standard U-Net, a number of variants were proposed. Firstly, the U-Net architecture was extended to 3D using 3D convolutions and poolings [68]. Secondly, residual connections can be added in convolution layers of the U-Net [69]. Finally, blocks of dense connections in the DenseNet [11] can replace the standard convolutions in the U-Net, which leads to the most recent Tiramisu Net [70].

2.3 Weakly Supervised Learning

2.3.1 Overview

Supervised learning methods are based on well-established datasets and the models trained on these datasets have been shown to be able to make accurate predictions. Here, a well-established dataset

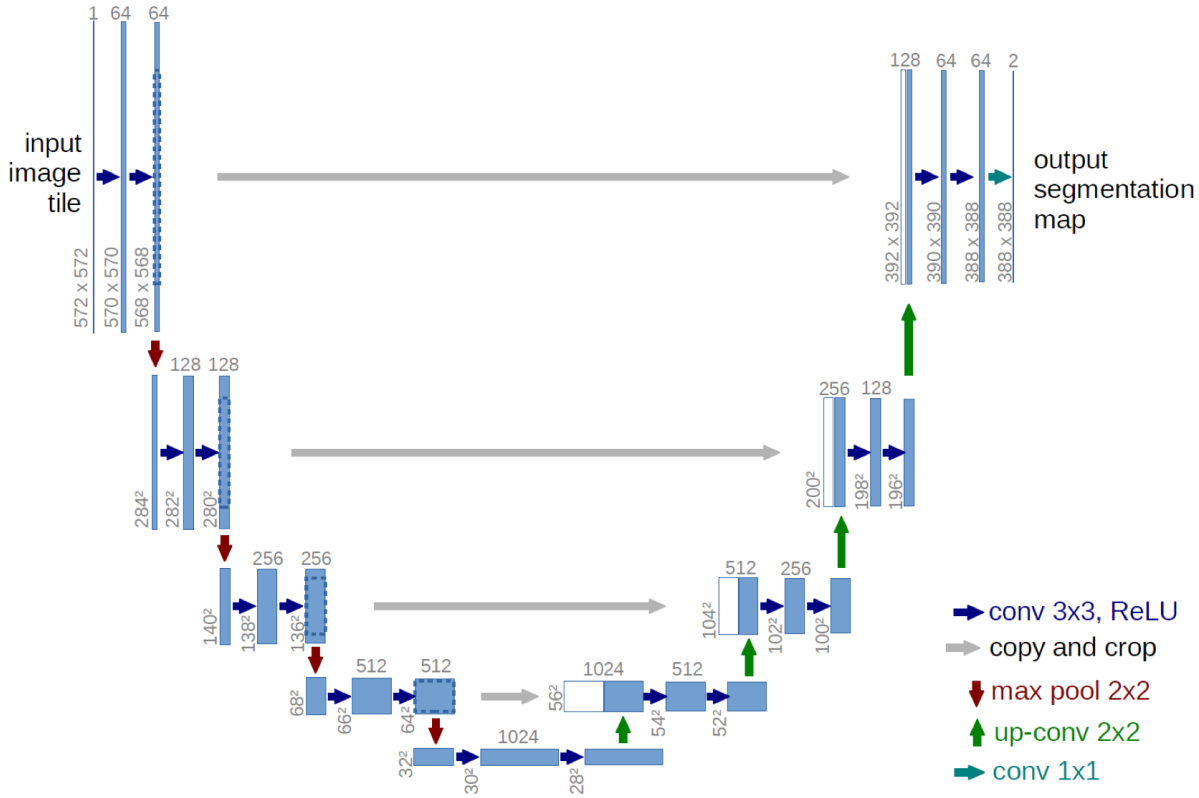


Figure 2.14: The architecture of the U-Net. The figure is from [6].

means each instance in the dataset has a ground truth label. However, it is usually difficult to establish such a dataset because of two major challenges: 1) Preparing labels for all instances is expensive, especially when the dataset is large; 2) Sometimes ground truth labels are not available and annotations by human experts are used instead. As such, disagreement is likely to occur among human experts. A dataset without complete ground truth labels is known as a weak dataset.

There are three types of weaknesses regarding instance labels. First, the labels can be *incomplete*, which means labels are only available for a subset of data. Second, the labels can be *inexact*, which means labels available are not as exact as expected, e.g. only coarse-grained labels are available. For instance, labels for pixels (or voxels) are needed to address the image segmentation problems. However, it is too expensive to annotate images at pixel (or voxel) level. Instead, labels at image level can be obtained. Third, the labels can be *inaccurate*. This may because of a few reasons including noise and the disagreement among human experts. In reality, these weaknesses happen individually or jointly, which makes it difficult to develop machine learning models.

2.3.2 Incomplete Supervision

There are two main strategies addressing incomplete supervision problems, namely active learning [71] and semi-supervised learning [72,73]. The difference between them lies in the human interaction. Human interactions are involved in active learning methods while semi-supervised learning methods are not based on human interaction.

Active Learning: A typical active learning approach works as follows [71]: First, build a model based on the limited data with labels; Second, query a label with human experts for an unlabelled data instance and retrain the model. Repeating the second step results in a good model. To minimize human interactions, the target is to raise minimum queries. Therefore, the key of active learning methods is selecting unlabelled data instances, which contribute the most to improve the model performance. Obviously, extra human resource is essential to active learning methods but it is not always available in practice.

Semi-supervised Learning: Semi-supervised learning [72, 73] is an alternative, which can make use of unlabelled data. It does not require extra human resource. The fact is that all the data instances are collected from the same source so that they obey the same distribution. Based on this fact, it is assumed that similar data instances have similar labels. The similarity can be measured by the distance between data instances. The distance is defined on the manifold representing the data distribution. There are four classic categories of methods of semi-supervised learning.

- Generative methods [74, 75]. These methods assume all the data instances are generated from the same model, e.g. Gaussian mixture model (GMM). Therefore, both labelled and unlabelled data are used to estimate the parameters of the model. The key to the generative methods is the accurate determination of models.

- Graph-based methods [76, 77]. Given a dataset, all the data instances are used to construct a graph. Each node in the graph corresponds to a data instance. The strength of an edge between two nodes represents their similarity. As such, labels of unlabelled data instances can be inferred via label propagation algorithms [76]. The implementation of algorithms is based on matrix computation, which is not efficient in terms of memory and computation.

- SVM-based methods [78–81]. The classic SVM can be extended to deal with unlabelled data. It is assumed that there is a low-density separation between data of different classes. Semi-supervised SVM aims to find hyper-planes which separates data instances through regions with less dense instances.

- Disagreement-based methods [82–84]. Multiple learners are employed by disagreement-based methods while other types of methods rely on single learners. These methods use the disagreement among learners to improve the performance of each learner. Therefore, the disagreement is the key to this type of methods. The disagreement is derived from the difference from the data and/or learners. Specifically, if a data instance has multiple attribute sets and each of them sufficiently represents the dataset and they are conditionally independent, then learners can learn different views from different attribute sets [82]. Alternatively, different learners can be employed [83]. In the learning process, each learner picks up unlabelled instances on which it has high confidence in terms of label assignment. Pseudo labels are then assigned to these instances and other learners will regard them as labelled instances in the next iteration. This means learners making different predictions on unlabelled instances teach and learn from each other until they make same predictions. This learning process promotes the performance of each learner.

Self-supervised Learning: Self-supervised learning is a generic learning strategy, which handcrafts supervised learning on unlabelled data to promote the learning accuracy on labelled data [4, 85]. Unlike the semi-supervised learning methods, in which unsupervised learning methods are usually involved, self-supervised learning handcrafts labels for the unlabelled data instances, enabling supervised learning on them. The handcrafted supervised learning approach learns critical features from the unlabelled data. The learned features are then transferred to models to be trained on the labelled data, which improves the learning efficiency and effectiveness.

The deep neural network (DNN) is a good model to work with self-supervised learning strategy. Specifically, training a DNN requires a large set of data with labels, which is expensive to obtain; otherwise the trained DNN cannot be generalized well. However, the majority of data instances are not labelled in practice. Training an extra DNN on the unlabelled data with handcrafted labels results

in representative features of the unlabelled data [5, 86, 87]. The learned weights of this extra DNN can be used to initialize the DNN to be trained on labelled data, which inherits the learned features of unlabelled data. The initialization enables the DNN to learn from the small dataset with labels. In addition, the initialization speeds up the subsequent training process.

Therefore, handcrafting labels for unlabelled data is the key to self-supervised learning. If the data consists of multiple modalities, handcrafting labels is fairly easy. For instance, videos can be viewed as image sequences. The temporal information is a good option to be used as self-supervised labels [88–90]. In terms of data with single modalities, self-supvervision labels may also be available. For instance, predicting local context of a static image can be regarded as self-supervised feature learning [5]. In this case, it is important to ensure the self-supervised DNN does learn semantic features of images instead of trivial features. If the local context to be predicted has a fixed position in all unlabelled images, the DNN is likely to only focus on the context around the fixed position.

2.3.3 Inexact Supervision

A dataset usually has some annotations. However, the annotated labels may not be as exact as needed. For instance, a dataset of images is labelled at image level while pixel (or voxel) level annotations are desired. This means there are only coarse-grained labels available. Therefore, the task in this scenario is to predict coarse-grained labels. Formally, a group of instances $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,m_i}\}$ comprise an instance bag $\mathbf{X}_i$. m_i is the number of instances. Labels of individual instances $\mathbf{x}_{i,j}, j = 1, 2, \ldots, m_i$ are unknown. If there is one instance $\mathbf{x}_{i,p}, p \in \{1, 2, \ldots, m_i\}$ in the bag is positive, then the bag is positive; otherwise the bag is negative. The target is learning a mapping $f : \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\} \mapsto \{-1, +1\}$. m is the number of bags in the dataset. This type of learning is called multi-instance learning [91].

In multi-instance learning, the proposed algorithms focus on learning at both bag level and instance level. Learning at bag level is similar to supervised learning. Therefore, existing supervised learning methods can be adapted to address the multi-instance learning problems. For example, the classic SVM approach can be extended as mi-SVM [92]. Ensemble methods including boosting and RF

can be adapted as MIS-Boost [93] and MIForests [94]. Learning at instance level, unsupervised learning methods are more suitable since labels for bag instances are not assigned [95]. Specifically, unsupervised methods can be applied for identifying distinctive instances, which make bags positive. This is based on the assumption that there must be at least one positive instances in positive bags.

In terms of applications, multi-instance learning is widely used, particularly where instances reside in high dimensional spaces as in the case of images [96–98]. More precisely, a high-dimensional instance could be regarded as sub-instances in lower dimension. For instance, an image can be regarded as a group of patches. Sub-instances of interest can be gathered in a bag, representing the instance. Different bags are likely to contain different numbers of sub-instances. Based on this setting, supervised learning algorithms are applicable to distinguish bags in different types. For the image example, distinguishing bag types means image classification, which is based on bags of patches.

2.3.4 Inaccurate Supervision

The basic idea to address problems of inaccurate labels is identifying these labels and correcting them [99]. Identifying an inaccurate label usually requires multiple experts to assign labels for the instance. Then the ground truth label of this instance can be inferred from multiple labels. Crowdsourcing [100] is a typical approach of collecting multiple labels for each instance. However, it is not always applicable. For instance, only human medical experts are able to interpret medical images. It is too expensive to employ a large cohort of medical experts. Ideally, a small number of data instances can be labelled by multiple experts. As such, a consensus label can be achieved based on multiple labels via ensemble techniques [101] Alternatively, training a model based on each expert's labelling results in multiple weak models [102]. A strong and robust model can be built by combining these weak models.

2.4 Summary

In the following chapters, several supervised and weakly supervised learning algorithms are used to develop novel approaches to solve medical image classification and segmentation problems. More precisely, in Chapter 3, a framework based on random forests is proposed to segment WMLs on CT images. In Chapters 4, 5, 6, CNNs are extensively used to develop methods for segmentation and classification problems. Finally, in Chapter 7, the SVD identification task is addressed based on a multi-instance boosting algorithm.

Chapter 3

Quantification of Cerebral Leukoaraiosis on CT

The work in this chapter is based on the following papers:

- L. Chen, A. Jones, G. Mair, R. Patel, A. Gontsarova, J. Ganesalingam, N. Math, A.C. Dawson, A. Basaam, D. Cohen, A. Mehta, J. Wardlaw, D. Rueckert, and P. Bentley, "Rapid automated quantification of cerebral leukoaraiosis on CT," *Radiology*, vol. 288, no. 2, pp. 573–581, 2018.
- O. Maier, B.H. Menze, J. von der Gablentz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, and others, "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Medical Image Analysis*, vol. 35, pp. 250–269, 2017.
- L. Chen, P. Bentley, and D. Rueckert, "A novel framework for sub-acute stroke lesion segmentation based on random forest," *Ischemic Stroke Lesion Segmentation Workshop*, 2015.

3.1 Introduction

Cerebral SVD – a major cause of age-related physical and cognitive morbidity – is most sensitively detected by FLAIR-MRI [28], typically as leukoaraiosis, i.e. WMLs, and lacunar infarcts. In practice,

WMLs are most commonly observed on CT images [103], rather than MR images, because of scanner availability and accessibility considerations in target populations. In acute stroke and traumatic head injury, CT is the first-line imaging modality of choice [104]; yet WML burden is an important variable, being a prognostic marker of functional outcome [39, 105, 106] and hemorrhagic transformation of ischemia [39, 107, 108]. For dementia, even though MRI is well-recognised to be superior in contributing towards diagnosis, hospital audits suggest that CT is used exclusively in the majority of cases [109–111].

Assessment of cerebral WML on CT images, is more challenging than using MRI, because signal characteristics of WML (hypoattenuation) are less distinctive relative to background white matter (WM) on CT images [22]. Moreover, sensitivity of CT decreases with smaller WML volumes [22, 112], and varies between brain regions [22]. Studies measuring inter-rater reliability of expert-based WML ratings show poorer agreement using CT than MRI [112, 113] (kappa coefficients from 0.5 to 0.6 for CT, versus 0.7 to 0.8 for MRI [22, 23]). Furthermore, WML scoring systems typically allow for only a small number of ordinal ratings (4–6 [114]), and use visual criteria (e.g. restricted to periventricular regions versus extending to cortex) that are imprecise, and do not convert directly to an estimate of total WML load [113]. As such, visual estimates of WML severity, although providing valuable prognostic information [39], have limited sensitivity as diagnostic markers, for monitoring disease progression, or in research.

In this study, we propose a novel framework to delineate WMLs on CT images and validate this method comprehensively, comparing the automated output with expert delineations on CT and MRI (i.e. gold standard), and ratings in about 1000 stroke patients, using images originating from a wide range of scanner types, thus reflecting typical populations that the technique is likely to be used in.

3.2 Methods

3.2.1 Study Populations

Since one of the primary applications for automated WML estimation is prognostication of acute ischemic stroke, the study focuses on this patient population. The cohorts (Figure 3.1) comprise : 1) all acute ischemic stroke patients presenting to ICL Hyperacute Stroke Unit between 2010 and 2014 who subsequently received thrombolysis treatment (ICL-thrombolysed cohort); 2) all acute ischemic stroke patients from ICL in the same time-period who underwent both CT and MRI within 1 week of each other (ICL CT-MRI cohort), excluding ICL-thrombolysed subjects; 3) a random sample of patients ($N = 200$) recruited to the IST-3 cohort [115], from which patients with obvious extensive acute ischemic changes are first excluded. This subset therefore is more typical of patients who might also present to a cognitive impairment clinic.

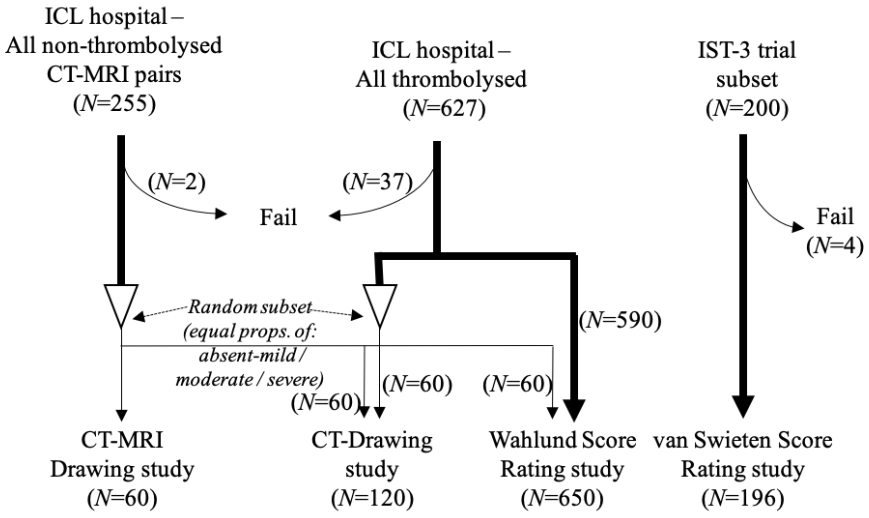


Figure 3.1: Flow chart of the cohorts involved in this study. In the cohort of ICL non-thrombolysed CT-MRI pairs, there was class imbalance so random subsets were used.

Testing of the automated WML quantification method is assessed by comparison with experts: 1) drawings of WML outlines on CT images and co-registered FLAIR-MR images (the latter is considered to be a ground truth), and 2) ratings using two conventional ordinal qualitative WML scoring systems [22, 23]. The Wahlund and van Swieten scoring systems are the most widely used systems for WML ratings. For the drawing study, 60 CT images are selected randomly from the ICL-thrombolysed cohort, and 60 from the ICL CT-MRI cohort, whilst ensuring that there are equal pro-

portions of absent/mild, moderate and severe SVD (based upon expert ratings). For the ratings study, ratings are obtained on all subjects from ICL-thrombolysed cohort, CT-MRI pairs and the IST-3 subsets. Table 3.1 describes subject characteristics, including imaging features, for each study.

Table 3.1: Sample characteristics of four validation studies. N denotes the numbers able to be processed by automated WML quantification method (i.e. excluding image processing failures). Atrophy studies were using atrophy grading system described in [1]. Other lesions include hydrocephalus, arachnoid cyst, meningioma, aneurysm, haemorrhage. In the IST-3, patients with acute ischemic parenchymal changes were excluded in advance.

	Drawing volume studies		Ordinal rating studies	
	CT only	CT-MRI pairs	Wahlund Score	van Swieten Score
N	120	60	650	196
Population description	Random selection of patients presenting to acute stroke ward; equal proportions of SVD severity: absent-mild/moderate/severe		All, unselected thrombolysed patients (+ CT-MRI pairs cohort)	Random selection of participants from, thrombolysis IST-3
Age (median, interquartile range (IQR))	76 (66-85)	76 (67-84)	75 (63-82)	82 (77-86)
Male (%)	52	58	54	45
CT features:				
acute parenchymal ischemia (%)	19	22	36	0
old infarcts (%)	38	38	42	59
central atrophy (%)	72	75	67	87
peripheral atrophy (%)	82	87	75	85
other lesions (%)	6	8	5	0
Expert raters (N):				
pool number	3	3	6	13
per scan	3	3	3	3

CT images used from ICL were derived from two types of CT scanners (GE, Siemens); comprised a range of slice thicknesses (voxel resolutions: about $0.4 \times 0.4 \times [1-7]$ mm), that in 70% of cases differed between the top- and bottom-halves of the brain (i.e. two image files per patient); and in the remainder were uniform volumetric images. The IST-3 cohort CT images comprised an even more heterogeneous set (details provided in original report [115]). The study was ethically approved by the ICL Joint Research Compliance Office.

3.2.2 Expert Drawings and Ratings

Experts are neuroradiologists or stroke physicians with more than 5 years of regular stroke experience. Those who performed testing drawings or ratings of WML are different to those who contributed to model training. Experts were trained in WML rating scores and/or digital lesion drawings prior to their assessments. Digital drawings (see Column 3 in Figure 3.4) were performed using the MRI-Cron software[1]. FLAIR-MR images were also annotated for WML, after first being aligned with each patient's contemporaneous CT [116], so as to minimise CT/MRI differences in WML appearances caused by variations in slice orientation. CT WML ratings used either the Wahlund [22] or van Swieten [23] scoring systems, reflecting 4 or 3 grades of WML severity, respectively. For the Wahlund system, experts recorded the median WML score across frontal, parieto-occipital and temporal regions [22]. For the van Swieten system, anterior and posterior scores [23] (3 grades each) were averaged and rounded. CT drawings and ratings were performed by 3 experts for each case, drawn from a pool of 3–13 for each experiment, allowing a consensus to be deduced for WML volume and rating score (mean and median respectively). Comparisons between each combination of rater pairs was performed to identify any experts who differed significantly ($p < 0.05$) in their performance.

3.2.3 Automated SVD Quantification

Overview: We propose to segment cerebral WMLs, which is leukoaraiosis including areas with lacunar infarcts [103], on CT images using RF [44]. Specifically, 2D image patches across multiple scales are extracted from CT images, WMLs on which have been manually annotated. The patch extraction is guided by a prior mask defining the ROIs of WMLs. An RF model is then trained based on the patches, classifying if the central pixel in the patch represents a WML.

Training Data: There are 90 representative CT slices, which are used for model development. These slices are selected from 50 subjects showing moderate or severe WMLs. The subjects involved are patients suffering from acute ischemic stroke (less than 4.5 hours from symptom onset). The scans used in training is from a separate stroke centre (Northwick Park Hospital).

[1] www.mccauslandcenter.sc.edu/crnl/mricron/

The 90 slices are randomly splitted into training and validation sets, which consist of 70 and 20 slices, respectively. Training and validation slices are from separate subjects.

Patch Extraction: Patches are extracted under guidance of a mask, which defines the ROIs of WMLs. This mask is achieved by Chen et al. [117]. More precisely, 277 FLAIR images with moderate or severe WMLs are collected and the lesions are annotated by experts. These FLAIR images are then registered to a common space, along with annotated lesions. As a result, an average lesion mask is obtained showing approximate probability of lesion occurrence at each voxel. Given any unseen CT image, a template in the common space can be registered to its native space. The transformation information is applicable to the lesion mask. Therefore, a lesion mask in the native space is obtained. Thresholding the lesion mask in native space creates the corresponding binary lesion mask. Extracting patches within the binary lesion mask has two major advantages: 1) other types of lesions outside of the ROIs can be excluded naturally; and 2) applying this algorithm is much faster.

Within the lesion mask in a native space, patches across multiple scales are extracted, which is shown in Figure 3.2. This follows the intuition that image features aggregating from multiple scales improve the performance of models [9]. In terms of implementation, original CT images are blurred using multiple Gaussian kernels, which results in images at multiple scales. As such, patches extracted at the same position from images at multiple scales represent local features across multiple scales. In this work, we set the patch size as 15×15 pixels.

The image blurring has another advantage in this case. Since CT images tend to have low signal-to-noise ratio, blurring removes some noise. Particularly, if the noise is so heavy that brain structures are damaged, denoising could highlight the brain structures, which makes the image features more robust. Figure 3.3 shows such an example.

WML Segmentation: We propose the intensities of pixels in the multi-scale patches as features. These features are then used to train a standard RF model. There are k decision trees build in a RF. In this study, we set $k = 100$ according to the model performance on the validation set. The output of the trained forest is the approximate probability of lesion occurrence at the central pixel of the patch.

Applying the trained RF model to each pixel within the mask generates the approximate probability

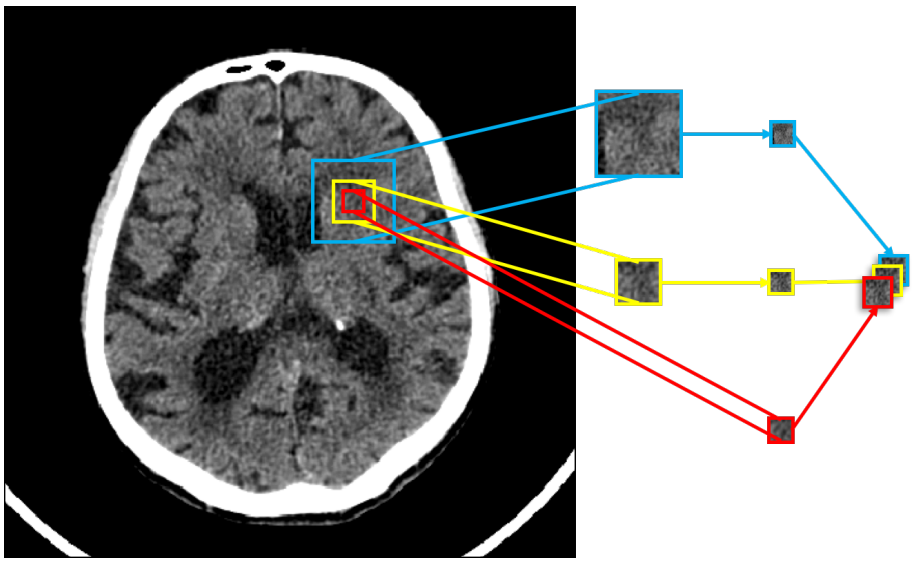
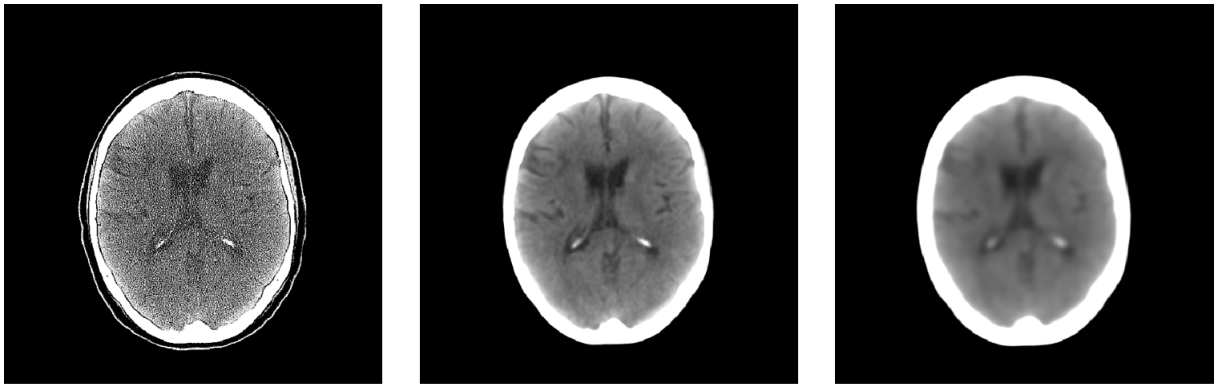


Figure 3.2: At each pixel of CT images, surrounding patches at multiple scales are extracted. Combining these patches results in the input to the following classifier.



(a) Original Image (b) Blurred Image $\sigma = 2$ (c) Blurred Image $\sigma = 4$

Figure 3.3: An example of CT image with heavy noise. Blurred by Gaussian multiple kernels, the image has higher signal-to-noise ratios. σ is the variance of the Gaussian filter.

map showing potential WMLs. Thresholding this probability map results in a binary WML map. The threshold of 0.2 is achieved based on the validation image set. Counting voxels in the binary lesion map and multiplying the voxel size result in the volume of WMLs.

Ordinal Rating Score Inference:

For comparison with ordinal rating scores, automated WML volumes are thresholded into ranks equivalent in number to the score systems [22, 23] used by experts (4 or 3). The thresholds are derived from the k-means clustering of estimated volumes.

3.2.4 Evaluation Methods

Drawings of WML on CT and MRI are compared for spatial similarity with automated segmentations using the patch-based evaluation of imaging similarity (PEIS) [118, 119], which is a metric similar to but more robust than the Dice score; and tested for group differences with the rank-sum test.

From expert drawings of cerebral WML on CT or MRI, total lesion volume is calculated, and correlated with automated WML volume, using Spearmans correlation. Comparisons of Spearman correlation coefficients are performed using an appropriate Fisher Z-Transformation [120].

Agreements between automated ratings versus expert ratings are assessed with linear weighted-kappa scores (κw), while comparisons between agreements are tested with validated bootstrap methods [121].

3.3 Results

3.3.1 Image Pre-processing

Image pre-processing consists of joining two halves of brain scans where applicable and registration from the template space to each native space. Image pre-processing failures occurred in 39 out of 882 hospital-derived CT images, and 4 out of 200 IST-3-derived CT images (3.98% total failure rate). Inspection of these cases identifies poor image quality, due to inappropriate intensity windowing, incomplete brain coverage, extensive movement, beam-hardening artefact, or extreme head tilt. Images in poor quality, which failed in pre-precessing, account for 42% of all failed cases (18 out of 43). Pre-processing takes 77.3 seconds in average (std of 25 seconds).

3.3.2 Lesion Segmentation

The median spatial similarity (PEIS) between automated WML delineations and expert MRI-WML drawings is 0.53 (IQR: 0.48–0.57) while the median PEIS between expert CT-WML and MRI-WML

drawings is 0.54 (IQR: 0.49–0.58). Strength of correlation between automated CT segmentations and expert drawings (CT or MRI) are not significantly influenced by age, sex, or co-existence of the following commonly-associated CT features: acute ischemic change, old infarct, central or peripheral atrophy, or other lesion. Therefore, the automated WML segmentations and expert CT-WML is not significantly different.

Figure 3.4 shows some visual examples of the WML segmentation, where the FLAIR-MR images are co-registered as reference. The last example in the figure has co-existing old territorial infarct and it is properly avoided by the automated algorithm.

Expert drawings take a median of 7.9 minutes per scan (range: 6.9–9.4), whereas automated method takes a median of 32 seconds (95% confidence interval (CI): 31–33 seconds) per scan. Correlation coefficients between rater pairs (CT-CT or CT-MRI) are not significantly different from one another.

3.3.3 Lesion Volume Estimation

Results displayed on Table 3.2 suggest that WML volumes estimated using automated method correlate closely with those derived from expert CT-drawings. The volume correlation between the automated estimation and censensus-Expert CT lesion volume is fairly strong ($r^2 = 0.71$) based on the 120 subjects involved in this test. It is tested that the expert CT-volumes are statistically different (Δr: $Z = 3.1$, $p < 0.01$). Correlation between expert CT-volumes themselves is higher ($r^2 = 0.85$). However, vertical lines in Figure 3.5(a) are long, which means that the range of expert CT-volumes per scan is wide. More precisely, the median expert estimate is 91% of mean estimate and the IQR is 55%–148%.

When we compare automated WML volumes with expert drawings of co-registered FLAIR-MRI, the correlation is stronger ($r^2 = 0.85$) and it is statistically significant (Δr: $Z = 3.8$; $p < 0.001$). The correlation between expert-CT versus expert-MRI WML volumes is also strong ($r^2 = 0.82$). The WML volumes of expert CT and MRI drawings are not from the same distribution. Therefore, automated WML volumes is comparable to expert-CT estimates in this sense. In addition, according to Figure 3.5(b), automated volumes of WML are more conservative than experts. Specifically, the

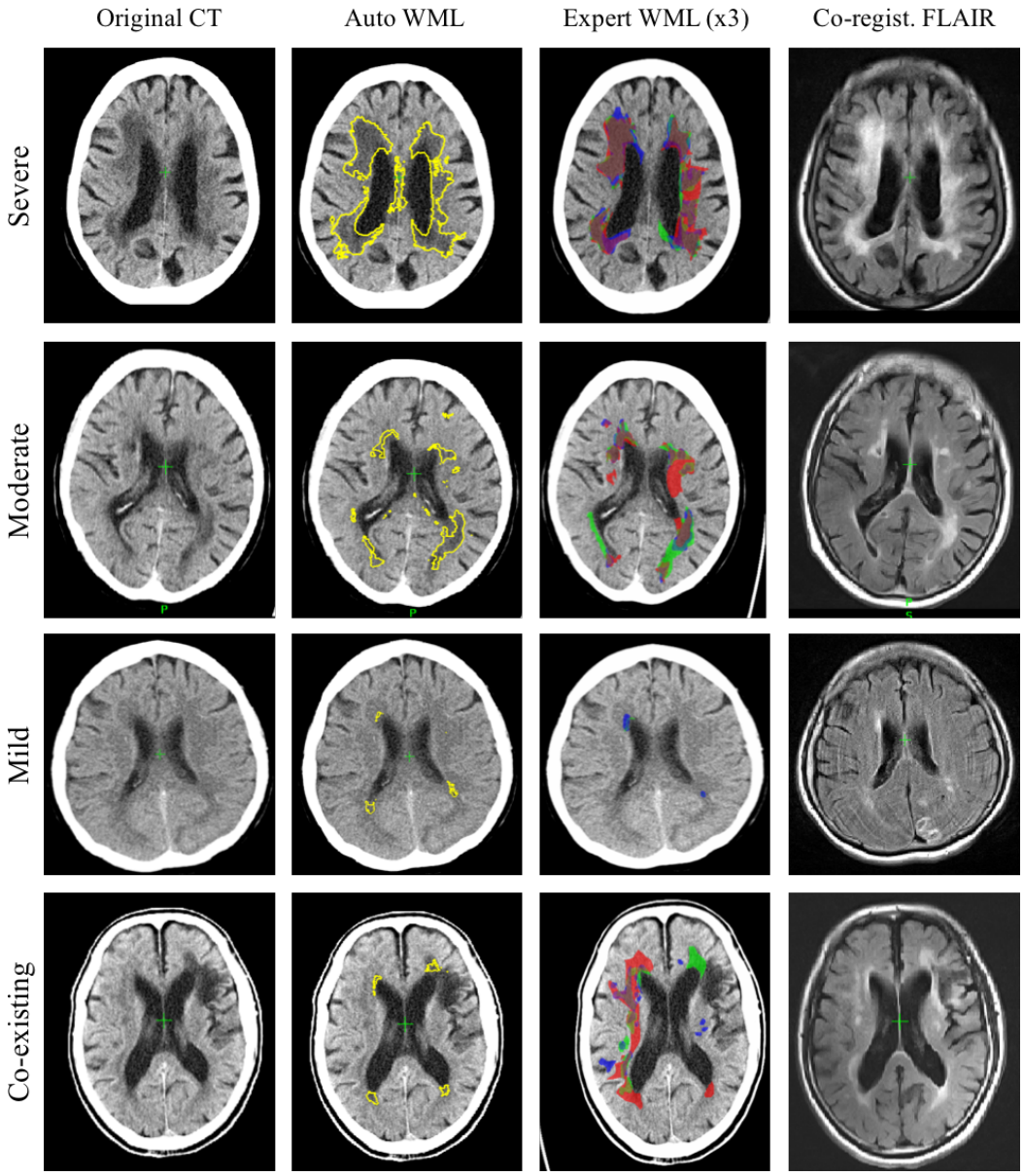


Figure 3.4: Examples of CT-WML delineations by automated method and Expert drawings (three colors represent specific experts annotations). The final column shows WML on co-registered FLAIRs, that are also delineated by experts (not shown here) and provided the ground truth.

automated WML volumes are lower than the lowest of three expert estimates in 43% ($p < 0.001$) cases and take 61% the value of mean expert CT-volumes (IQR: 40%–112%).

3.3.4 Ordinal Rating

Table 3.3 displays the agreements between automated ratings (i.e. thresholded WML-volume estimates) and individual experts ratings. In the Wahlund scoring system [22], agreement between

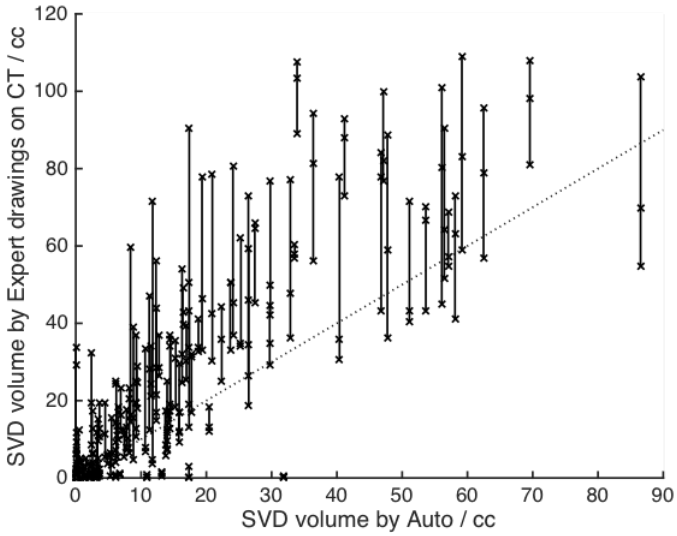
Table 3.2: Correlations between expert drawing and automated volumes. Range here refers to automated volumes vs individual expert drawing volumes. All correlations are significant at $p < 0.001$.

Study	Correlation of lesion volume between:	r^2	Range
CT only	automated volumes versus consensus-Expert CT lesion volumes (mean of 3)	0.710	0.645–0.713
	Expert CT drawings between themselves ($\times 3$)	0.845	0.813–0.867
CT-MRI pairs	automated volumes versus consensus-Expert MR lesion volumes (mean of 2)	0.850	0.823–0.833
	Expert CT drawings with Expert MRI drawings	0.819	0.767–0.856
	Expert MR drawings between each other ($\times 2$)	0.937	–

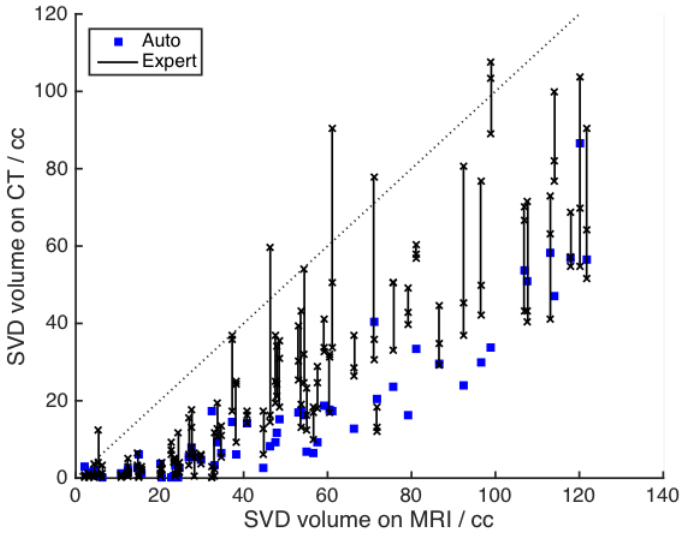
automated ratings and individual experts ratings is moderate (κw$= 0.529$). The ratings of experts are tested that they are not from the same distribution. Agreement between expert pairs is also moderate (κw$= 0.506$). In addition, agreement between automated ratings and expert consensus ratings is higher (κw$= 0.599$, $\Delta\kappa$w$p < 0.001$), which is also shown in Figure 3.6(a). Correlations of automated WML volume with expert ratings is also greater using consensus ($r^2 = 0.582$), than individual expert ratings ($r^2 = 0.506$, Δr: $Z = 2.05$, $p < 0.05$). Therefore, the automated ratings are comparable to experts ratings in the Wahlund scoring system.

Using the alternative van Swieten grading system [23], inter-expert agreements are higher (κw$=$ 0.665) than using the Wahlund system ($\Delta\kappa$w$p < 0.01$), and also higher than the agreement between automated method and individual experts (κw$= 0.571$; $\Delta\kappa$w$p < 0.05$). This because the van Swieten system has three grades, which is one less than the Wahlund system so there are less disagreement between experts. However, inter-expert agreement is not significantly different to the agreement between automated method and expert consensus. Correlations between automated WML volume and expert consensus van Swieten ratings ($r^2 = 0.629$) do not differ to that between automated and expert-consensus Wahlund ratings, and individual-expert van Swieten ratings. Therefore, the automated ratings are also comparable to experts' ratings in the van Swieten scoring system.

There are no clear boundaries, which can separate WML into different severity groups quantitatively. It is acceptable that the ratings by experts or other methods differ within 1 point. The proportion of cases in which automated rating is > 1 point different from expert consensus (i.e. strong disagreement) is 0.046, and 0.02, for Wahlund and van Swieten rating systems, respectively. Figure 3.6 shows



(a)



(b)

Figure 3.5: (a) Correlations of automated WML volumes with Expert drawings on CT. Each of three experts is indicated by a "×", with a connected line showing range of expert values. (b) Correlations of gold-standard WML volumes (expert drawings on FLAIR-MRI) with automated volumes (blue squares), and expert drawings on CT (each of 3 experts marked by "×"; range shown by vertical line). Dashed line of equality shown in each case, indicating that estimated WML volumes for any one patient tend be in order: automated WML < expert CT-WML < expert MRI-WML.

these outliers.

Inter-rater agreements between any particular expert pairs, using either rating system, do not differ significantly from one another. Time charts of raters (for Wahlund ratings) suggest that 30 scans took

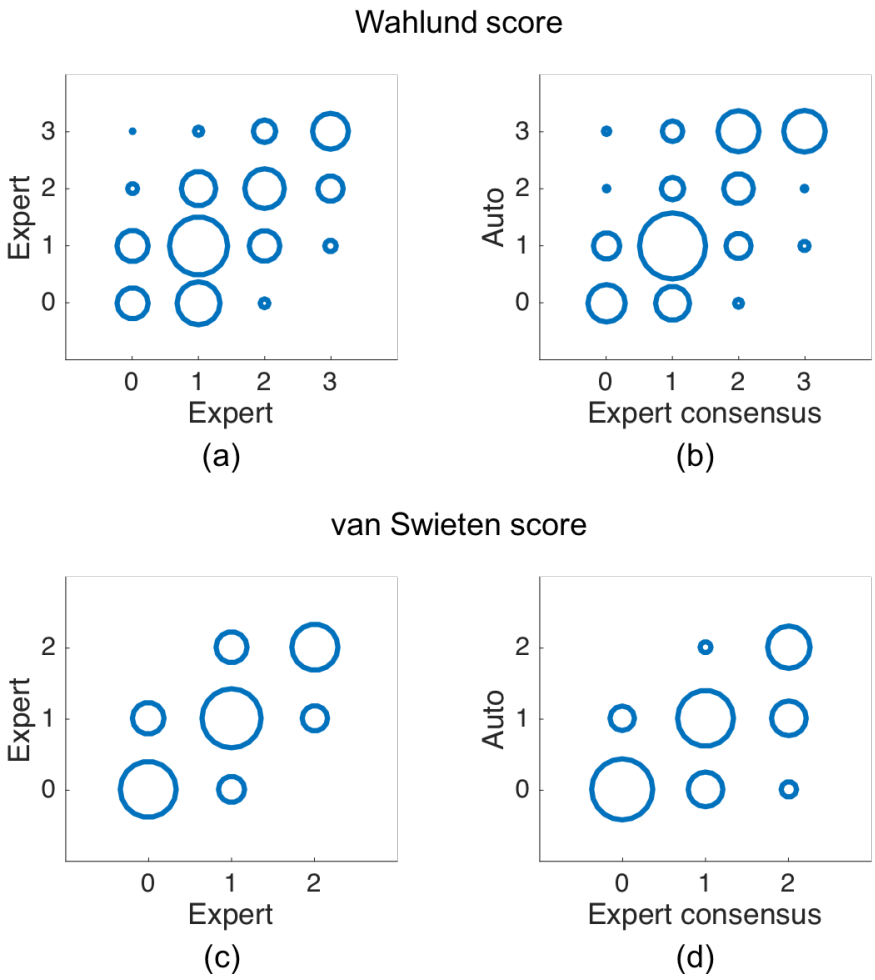


Figure 3.6: Agreement plots of expert-expert and automated-expert consensus for two CT-WML scoring systems. Automated score based upon thresholding of automated WML volumes.

about 45–60 minutes to rate, i.e. about 1.5 to 2 minutes each in total. The human rating process includes image-file selection, contrast adjustment, and judgements of three cerebral locations.

3.4 Discussion and Conclusion

We propose a novel framework, enabling accurate, fully-automated, and rapid quantification of cerebral leukoaraiosis (WML) on CT, in a large, multi-centre dataset. The automated method performs similarly to expert CT WML delineations in terms of lesion volume and spatial similarity, relative to a gold-standard of expert delineations of WM hyperintensities on co-registered T2-FLAIR [28]. Additionally, by converting automated WML volumes into ratings, agreements with experts' CT-WML visual ratings are similar to those comparing agreements between expert pairs themselves. In the

Table 3.3: Agreements and correlations between expert and automated scores or volumes.

Study	Agreement (κw) of SVD score ratings between:-	κw	Range
Wahlund Score (0-3)	Experts amongst themselves ($\times$6) [see Figure 3.6(a)]	0.506	0.473–0.552
	Auto versus Experts (individuals)	0.529	0.465–0.579
	Auto versus Expert (consensus) [see Figure 3.6(b)]	0.599	0.586–0.611
	Correlation of Expert SVD score rating and Auto volume	r^2	Range
	Expert individuals	0.506	0.462–0.549
	Expert consensus	0.582	–
van Swieten Score (0-4)	Agreement (κw) of SVD score ratings between:-	κw	Range
	Experts amongst themselves ($\times$3) [see Figure 3.6(c)]	0.665	0.648–0.674
	Auto versus Experts (individuals)	0.571	0.534–0.597
	Auto versus Expert (consensus) [see Figure 3.6(d)]	0.636	0.517–0.747
	Correlation of Expert SVD score rating and Auto volume	r^2	Range
	Expert individuals	0.571	0.522–0.614
	Expert consensus	0.629	–

largest cohort, agreement is greater for comparisons of automated versus expert consensus ratings, than versus individual expert ratings (or agreements between expert individuals themselves), which supports automated method, given that consensus opinions generally lie closer to the truth [122]. Images comprise a range of image resolutions, scanner qualities, and hospital origins, and are derived from centres separate to that which contributed training images, indicating the technique's robustness. Furthermore, accuracy of automated WML estimation is not hindered by common, co-existing hypoattenuating lesions e.g. acute or chronic ischemia, or atrophy.

At the same time, our study confirm previous findings that standard visual inspection methods for CT-WML estimation result in relatively modest interrater agreement: with kappa values of 0.5–0.6 being typical for common rating systems [22, 23, 112, 113]. This is also shown by the finding that expert CT delineations resulted in a wide range of estimated WML-volumes, even though they correlate strongly with each other ($r^2 = 0.85$). By contrast, the automated method always results in the same estimate of WML volume, once model parameters have been set. Importantly, the parameters of the model tested here do not alter, and are based upon an independent training dataset. Thus the automated method allows for a reduction in variable noise compared to existing WML scoring techniques, potentially enabling more reliable diagnostic and prognostic models to be developed.

A further asset of the automated method is that processing time averaged 109s including image pre-

processing, with the range being less than 3 minutes, which is similar to experts performing visual ratings. Considering that images originated from a number of centres, and CT scanners, this performance metric suggests that the automated method could be used widely in emergency rooms for rapid estimation of background WML from CT. The technique's option of superimposing machine-identified WML (Figure 3.4) can provide extra physician reassurance regarding the algorithm's output, and assist imaging interpretation by clinicians who are not so experienced in this.

Notwithstanding the automated method's advantages, we also draw attention to its limitations. CT images can not be processed in approximately 4% of cases, that are only partially accountable by poor image-quality issues. Additionally, among images that are processed, significant errors are made (> 1 point from consensus rating) in approximately 4%. Although small discrepancies with consensus are made in approximately 30% of cases, it is important to note that expert ratings are based upon judging categorical features (e.g. focal versus confluent lesions; extension to cortex or not) that are not directly proportional to lesion volume. Hence a better judge of automated method's accuracy is measuring discrepancy of automated estimates from volumes of expert drawings. In this regard, while automated-versus-expert drawing correlations are strong, there is also a consistent underestimation of automated WML volume relative to expert volumes (Figure 3.5). Threshold on the lesion probability map has a significant impact on the final binary lesion maps. According to the validation set, the optimal threshold was set as 0.2. However, we drew the ROC curve (Figure 3.7) based on the testing set with expert annotations on CT images and found a smaller threshold (i.e. 0.1) can result in better estimation. To this end, the threshold could be set as 0.1 in the future. Alternatively, drawings on MRI can be mapped to CT and used for model development. Furthermore, the fact that automated WML segmentations spatial similarity to MRI-WML is not significantly different to experts CT annotations, despite the former being smaller, indicates that the additional areas annotated by experts are not as accurate as the core areas identified by both automated method and expert.

The main reason for quantifying WML on CT, rather than MRI, is practicality. CT is the principle neuroimaging modality for emergencies such as acute stroke [104], and head trauma; and is often the sole imaging technique for investigation of dementia [109–111]. CT-analytic approaches have been developed recently to try to delineate chronic [123], and acute ischemia [124], as well as to predict hemorrhagic transformation after ischemic stroke [125]. One promising application for WML

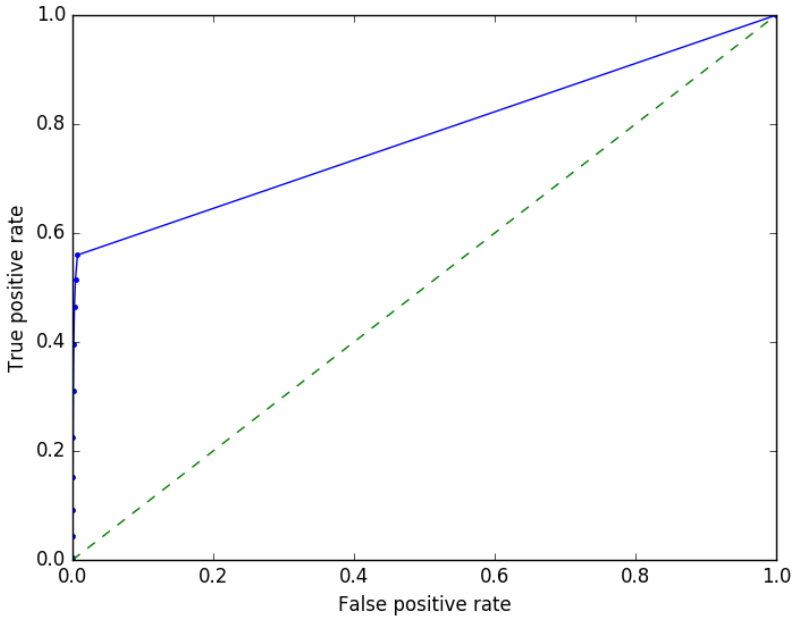


Figure 3.7: The ROC curve of the proposed model performance based on the testing CT image set with expert drawings. It shows that the optimal threshold is 0.1.

quantification is treatment-selection for acute ischemic stroke, given that cerebral WML load predicts poor functional outcome [39, 105] and ICH transformation [107, 108]. Currently this CT imaging predictor and others, e.g. acute ischemia extent, have not been found to interact with thrombolysis (or thrombectomy) treatment in their association with ICH so they are not recommended for hyper-acute treatment stratification [39, 126]. Since automated CT feature extraction, as presented here for WML, offers a reduction in variable noise relative to expert ratings, it would be interesting to explore whether such methods can identify treatment-specific ICH or functional outcomes. A related application would be to see if CT WML quantification could be used to predict anticoagulant-associated ICH [127] or hematoma growth and early deterioration after primary ICH [128]. More generally, WML quantification may be important in diagnosing, grading and monitoring vascular dementia (and possibly other types of dementia); and for prognosis after head injury [106].

In summary, automated CT-WML quantification enables reliable parameterization of a common biomarker of cerebral SVD. Clinical decision-making or research, in which WMLs are relevant, and where CT is the predominant imaging modality, may benefit from the method more than existing observer-dependent visual ratings.

Chapter 4

DRINet for Medical Image Segmentation

The work in this chapter is based on:

- L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "DRINet for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2018.

4.1 Introduction

Chapter 3 shows an example of medical image segmentation. Significant progress has been achieved in the field of medical image analysis in recent years due to the advent of CNNs [129]. Among the different approaches that use CNNs for medical image segmentation, the U-Net architecture [6] and its 3D extension [68] are widely used because of their flexible architectures. In the first part of the U-Net architecture (analysis path), deep features are learned while the second part of the U-Net architecture (synthesis path) performs segmentation based on these learned features. Training the two parts of the network in an end-to-end fashion yields good segmentation results. As the number of features in the first part of network is reduced because of convolutions and poolings, skip connections are used to allow dense feature maps from the analysis path to propagate to the corresponding layers in the synthesis part of the network, which improves the performance significantly.

However, the limitation of the U-Net architecture is its scalability. Specifically, deeper networks learn more representative features and result in better performance. The most straightforward way to make the U-Net architecture deeper is to add more layers. Adding more layers to the network enlarges the parameter space, which allows the network to learn more representative features. However, this also increases the difficulties in training the network because gradients are likely to vanish during training. Therefore, the challenge is to make the network wider and deeper without gradient vanishing.

In computer vision, the state-of-the-art CNN architectures (until 2017) include the densely connected convolutional network (DenseNet) [11, 130] and the Inception-ResNet [15]. The DenseNet approach consists of a number of dense blocks with pooling layers between them to reduce the size of the feature maps. Within each dense block, layers are directly connected with all of their preceding layers, which is implemented via concatenation of feature maps in subsequent layers. This dense architecture has a number of advantages: Firstly, the concatenation of feature maps enables deep supervision so that gradients are propagated more easily to preceding layers, which makes the network training easier. Secondly, bottleneck layers (convolution layers with 1-by-1 kernels) are used to control the growth rate of parameters in the network. Finally, in the DenseNet architecture the final classifying layer uses features from all layers (instead of only features from the last layer as in standard CNN approaches), leading to improved classification performance.

The Inception [9] is a CNN architecture which uses the Inception modules and allows for very deep networks. The main purpose of the Inception modules are: 1) to increase the depth and width of networks without adding more parameters; and 2) to achieve multi-scale features for processing. This is achieved by carefully designing structures of the Inception modules. The latest version of the Inception architecture [15] also uses residual connections, i.e. Inception-ResNet. Figure 4.1 shows an overview of the Inception-ResNet: a stem convolution block, stacks of Inception and reduction blocks, and the classifier. The stem block consists of a number of standard convolution and pooling layers, reducing the size of feature maps in lower layers (the ones close to the input). This aims to be memory efficient in training but it is not strictly necessary. Each Inception block consists of a number of Inception modules. The Reduction blocks are Inception modules with dimension reduction. An Inception module consists of a number of branches of convolution layers. In each branch, a bottleneck layer reduces the number of feature maps. The feature maps are then processed by convolution layers

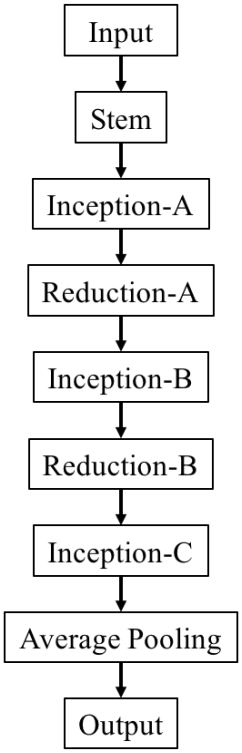


Figure 4.1: The overall schema of the Inception-ResNet [15]. The whole architecture consists of some Inception and Reduction blocks. Each block contains a number of modules. The detailed structures in different blocks vary slightly.

with different sizes of kernels in different branches. The outputs of all branches are finally aggregated as the output of the Inception module.

Inspired by the DenseNet and the Inception-ResNet, we propose an architecture consisting of dense connection blocks, residual Inception blocks, and unpooling blocks. We term this architecture Dense-Res-Inception Net (DRINet). We apply the proposed DRINet architecture for three challenging clinical segmentation problems, namely multi-class segmentation of brain CSF in CT images, abdominal multi-organ segmentation in CT images, and brain tumour segmentation (BraTS) in multi-modal MR images. They are based on clinical datasets and particularly the last problem is based on a publicly benchmark dataset. Our main contributions are: 1) a novel combination of the dense connections with the Inception structure to address segmentation problems. The use of dense connection blocks, residual Inception blocks, and the unpooling blocks achieves high performance while maintaining computational efficiency; 2) easy and flexible implementation of the proposed network architecture; 3) state-of-the-art segmentation performance for challenging image segmentation tasks.

4.2 Related Work

The basic CNN architecture for many semantic segmentation problems is the FCN, shown in Figure 4.2(a), which consists of cascaded convolution, pooling, and deconvolution layers. The convolution and pooling layers form the analysis path while the convolution and deconvolution layers form the synthesis path. The analysis path and the synthesis path are usually symmetric.

The U-Net (Figure 4.2(b)) is the FCN with skip layers between layers in analysis path and synthesis path. The skip layers are implemented via concatenations and they allow deep supervision for the network. Therefore, the skip layers improve the network performance. In addition, residual connections can be used in the U-Net, which results in the Res-U-Net (Figure 4.2(3)). In the Res-U-Net, the residual learning is implemented using the bottleneck building blocks with residual connections, which were used in the ResNet-50/101/152 architectures [10].

The DeepLab approach [14] involved atrous convolutions and poolings within the CNN architecture to solve segmentation problems, as well as CRF models for post processing. Based on the DeepLab architecture, Chen et al. [66] proposed the latest DeepLabV3 architecture. In DeepLabV3, a simple synthesis path is used. This synthesis path only consists of very few convolution layers, which is different from the synthesis path used in the FCN and the U-Net architectures. Skip connections are used to connect the analysis path and the synthesis path.

Finally, the DenseNet was extended in a fully convolutional fashion so that it fits segmentation tasks [70]. Specifically, an upsampling transition module was proposed in correspondence to the downsampling transition module in the original DenseNet. In addition, the macro architecture of the fully convolutional DenseNet is similar to the U-Net where skip connections are used.

The pyramid scene parsing network (PSPNet) [131] was proposed to solve the challenging scene parsing problem. In the scene parsing problem, prior knowledge could be incorporated in CNNs to improve performance. For example, cars are likely to be on the road while they should not be in the sky. Global context is required to incorporate these priors. The pyramid pooling module in the PSPNet investigate features in multiple levels, achieving the state-of-the-art performance.

4.3 DRINet

4.3.1 Overview

Figure 4.2(d) demonstrates our proposed DRINet architecture. Similar to the FCN, the DRINet has an analysis path and a synthesis path. Stacks of dense connection blocks, instead of standard convolution layers make up the analysis path, which is inspired by the DenseNet. The synthesis path consists of residual Inception blocks and unpooling blocks, which are inspired by the Res-Inception Net. To be more efficient in terms of memory, the DRINet has no skip connections. In this work, we show the DRINet architecture in 2D but it is straightforward to extend it to 3D.

4.3.2 Dense Connection Block

We employ convolutional dense connection blocks [11] in the analysis path, which are shown in Figure 4.3. Formally, let us assume x_l is the output of the l^{th} layer and $f(\cdot)$ is a convolution function followed by BN [132] and ReLU. In the standard convolution layer, we have:

$$x_{l+1} = f(x_l) \tag{4.1}$$

while in the dense connection block [11] we have

$$x_{l+1} = f(x_l) \circ x_l. \tag{4.2}$$

Here $\circ$ indicates concatenation.

The number of output channels from standard convolution layers are usually fixed and typically 64 or 128. As a result, it is expensive in terms of memory to concatenate the outputs of preceding convolution layers. In addition, the concatenation also leads to many redundant features. Therefore, Huang et al. [11] propose to heavily reduce the output size via 1×1 convolutions. As shown in Figure 4.3, within a dense connection block, the size of the output channel for each convolution layer k_i is typically small, e.g. 12 or 24 and this is commonly referred to as the growth rate of the network.

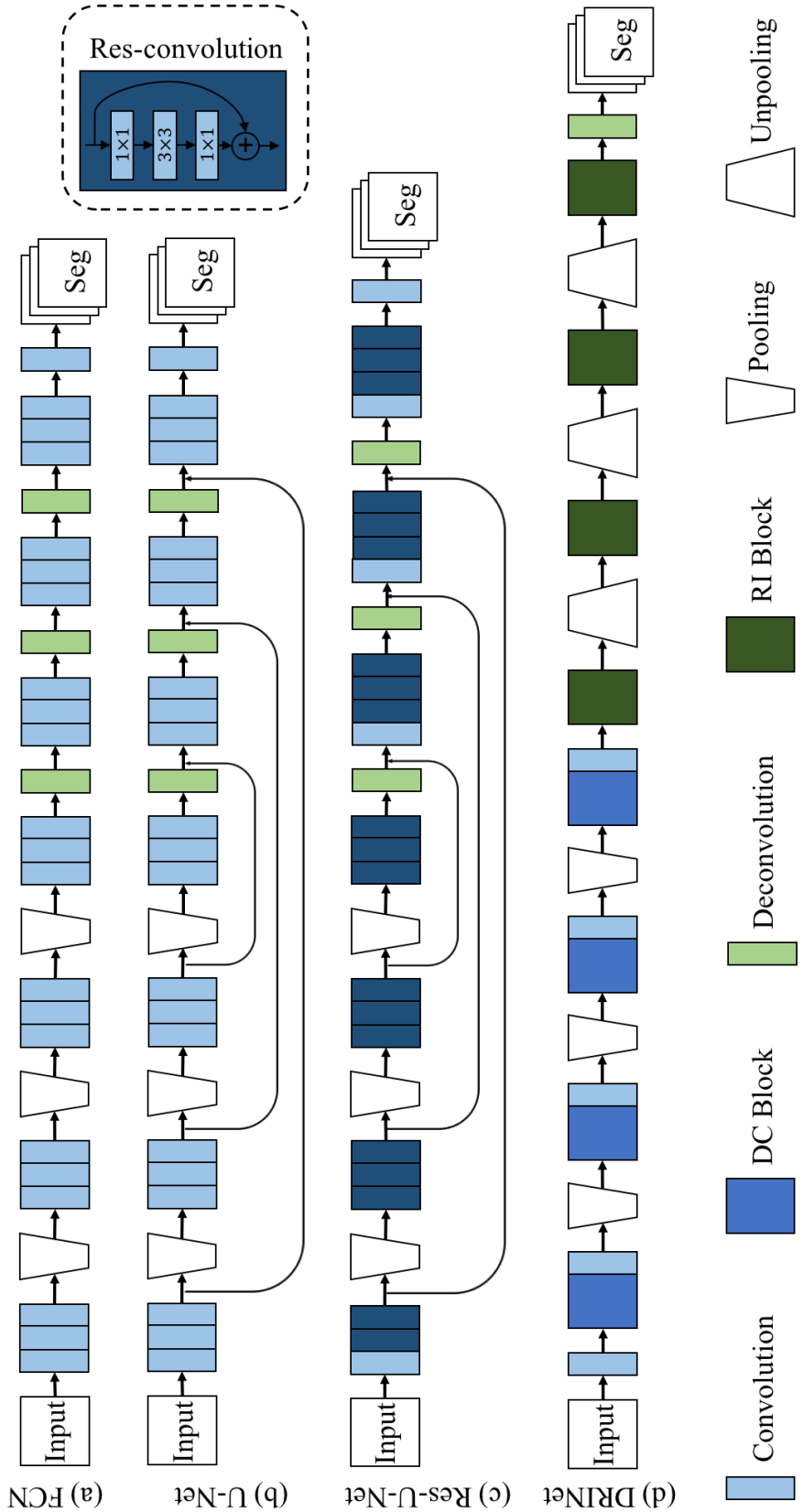


Figure 4.2: Overview of the FCN, the U-Net, the Res-U-Net and the DRINet. DC block and RI block represent the dense connection block and the residual Inception block. In the DRINet, the DC, RI, and unpooling blocks are depicted in Figure 4.3, 4.4, and 4.5, respectively. In the Res-U-Net, the residual convolution means the bottleneck building block used in the ResNet-50/101/152 [10].

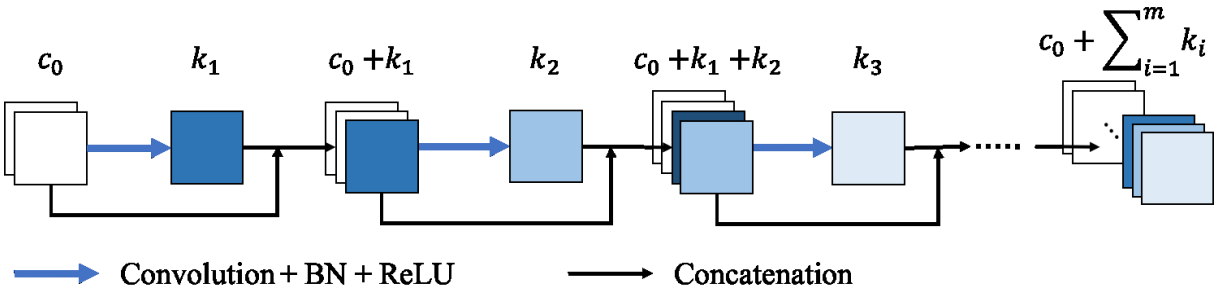


Figure 4.3: A dense connection block contains m convolution layers. The output channel number of each convolution layer k_i is the growth rate. The numbers (e.g. $c_0 + k_1$) above rectangles are the resulted number of channels in each layer. BN and ReLU apply on every convolution layer. The input and output of a convolution layer is concatenated so deep supervision is allowed.

Using dense connection blocks in the analysis path leads to three major advantages: 1) Gradient propagation through the network is more efficient. Conventionally, it is difficult to ensure that gradients backpropagate to lower layers in the network. Therefore, it is important to use dense connection blocks to alleviate the effect of vanishing gradients. 2) The input to the synthesis path consists of feature maps output from all preceding layers, instead of only the last layer. These feature maps lead to better segmentation results. 3) It is easy to use the growth rate to control the parameter space, resulting in good network performance.

4.3.3 Residual Inception Block

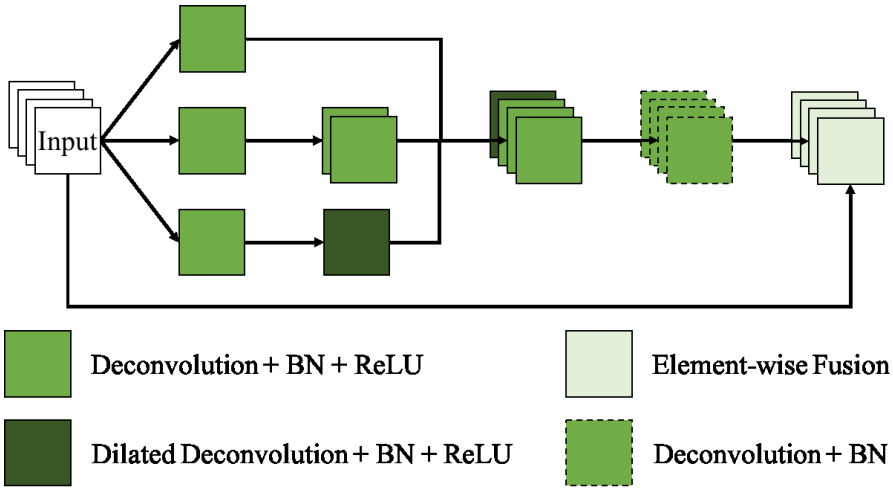


Figure 4.4: A residual Inception block is an Inception module with residual connections. An Inception module is a weighted combination of features maps from a few branches. Each branch process the input feature maps using deconvolutions with different kernel sizes.

In the synthesis path of the DRINet, we propose to use the residual Inception blocks, which is depicted

in Figure 4.4. Similar to the original Inception modules [9], the idea is to aggregate feature maps from different branches, where the input feature maps are convolved using kernels in different sizes. The residual connections make the learning easier since a residual inception block learns a function with reference to the input feature maps, instead of learning an unreferenced function.

In terms of the kernel sizes in convolutions, it is difficult to determine the optimal size for each convolution. In the FCN and the U-Net, the kernel size of convolutions is fixed as 3×3. In the inception module, convolutions of different kernel sizes are used in parallel. The weights can be learned in each inception module. In implementation, the feature maps are combined using concatenation and a deconvolution layer with 1×1 kernel learns the combination weights. The deconvolutions in the proposed Inception modules work the same as the convolutions. The purpose of this is to differentiate with convolutions in the analysis path in symbols.

Unlike the Inception Res-Net [15] having various Inception modules, we propose to use identical Inception blocks in the DRINet, which is easy to implement. We propose to aggregate feature maps convolved by three kernels, namely 1×1, 3×3, and 5×5. Inspired by the DeepLab [24], the deconvolution with a 5×5 kernel is replaced by a dilated deconvolution with a 3×3 kernel, which is more efficient in memory. To further limit the parameter space, a bottleneck deconvolution is used in each branch.

Formally, let $g(\cdot)$ denotes a deconvolution function followed by BN and ReLU and $g_b(\cdot)$ and $g_d(\cdot)$ represent bottleneck and dilated deconvolution respectively. As a result we obtain

$$x_{l+1} = g_b(g_b(x_l) \circ g(g_b(x_l)) \circ g_d(g_b(x_l))) + x_l. \tag{4.3}$$

4.3.4 Unpooling Block

We propose an unpooling block shown in Figure 4.5 to upsample the feature maps in the synthesis path. The unpooling block can be viewed as a mini Inception module, which combines upsampled feature maps from two branches. In each branch, the input feature maps are convolved using kernels in different sizes, namely 1×1 and 5×5. The resulting feature maps are then upsampled using

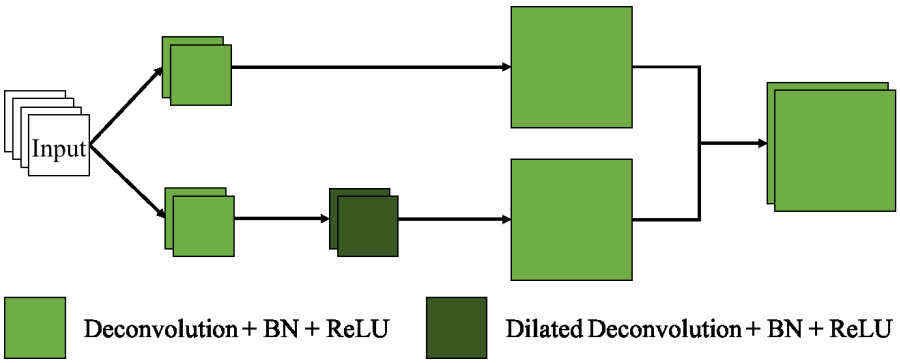


Figure 4.5: An unpooling block is a mini Inception module and it upsamples the input feature maps.

a deconvolution layer with stride 2. Again, the deconvolution with a 5×5 kernel is replaced by a dilated deconvolution with a 3×3 kernel in order to ensure memory efficiency. Also, to limit the parameter space, the input feature maps are firstly convolved by a bottleneck layer in each branch, which is similar to it in the residual Inception block. The combination of upsampled feature maps is achieved via concatenation. Formally, let $g_2(\cdot)$ denotes the deconvolution function with stride 2. The upsampled feature maps are therefore:

$$x_{l+1} = g_2(g_b(x_l)) \circ g_2(g_d(g_b(x_l))). \tag{4.4}$$

The major advantage of the proposed unpooling block is the aggregation of different upsampled feature maps. Specifically, simply upsampling the input feature maps using a deconvolution layer is likely to produce errors. For instance, a small error in the input feature maps is likely to be enlarged, which finally results in errors in the segmentation results. In contrast, convolving the input feature maps with different kernels leads to different intermediate feature maps. Upsampling these feature maps separately and combining them together reduce the effect of errors. This is the idea of ensemble.

4.3.5 Evaluation Metrics

In multi-class segmentation on brain CSF and abdominal organs, we use the well-known Dice coefficient as well as sensitivity (SE) and precision (PR) for evaluation. In evaluation in the BraTS challenge, we use the same metrics used in the challenge, namely the Dice coefficient, the SE, the specificity (SP), and the Hausdorff95 distance. The Hausdorff95 distance is a robust version of the

standard Hausdorff distance, which measures 95 quantile of the distance between two surfaces, instead of the maximum.

4.3.6 Implementation Details

In this work, we use cross-entropy as the loss function for all networks. We use the Adam method [133] for optimization with the following parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$. An initial learning rate of $1e-3$ is utilized. The weights are all initialised from a truncated normal distribution of standard deviation of 0.01. BN [132] layers are employed in all convolution and deconvolution layers except the last convolution/deconvolution layer. There are three convolution layers in each dense connection block and the kernel size is 3×3 with stride 1. There are three residual Inception modules in each residual Inception block. For the standard deconvolution layers in the residual Inception module, the kernel size is 3×3 and the stride is 1. All networks used in this chapter are implemented on the Tensorflow[1] platform.

4.4 Experiments and Results

4.4.1 CSF Segmentation in CT Images

Overview: Assessment of CSF volume, within ventricles and cortical sulci, is important for numerous neurological and neurosurgical applications. In many applications where rapid assessment is required (e.g. stroke), CT is preferred over MRI [134]. A common condition requiring the quantification of CSF is hydrocephalus (ventricular enlargement), a potentially life-threatening, but reversible condition; caused by a wide range of pathologies including hemorrhage, edema or tumours [135]. In these cases, CSF space quantification, especially comparison of ventricular to sulcal compartments, is important for distinguishing hydrocephalus from atrophy (due to age-related ischemia or degeneration) [38]. Standard quantification methods rely upon simple measurement of ventricular spans [136].

[1] https://www.tensorflow.org/

However, given the complex ventricular shape, these are imprecise, vary between observers and do not allow for accurate estimation of sulcal CSF [137].

The challenges for multi-class CSF segmentation in CT are three-fold: 1) clinical CT images are often acquired as stacks of 2D image slices with large slice thickness. Thus, each slice is usually separately analyzed, however the position of the patient's head is usually highly variable. Therefore, the CSF on each 2D image slice can vary significantly in terms of its configuration and shape; 2) patients often have background disease (e.g. old infarcts) which can have similar intensities to the CSF. It is not easy to distinguish between them as they are likely to mix together; 3) at the borders of different categories of CSF, segmentation errors often occur. Many existing methods [138–146] are not robust to these problems. To the best of our knowledge, this is the first attempt to solve the multi-class CSF segmentation problem in CT images.

Dataset: CT scans from 133 stroke patients were collected from two local hospitals. All clinical CT scans were collected retrospectively from local PACS databases and anonymized before performing research. Ethical approval was obtained from the ICL Joint Research Office. The scans were acquired on three types of CT scanners (GE, Siemens, and Toshiba). The thicknesses of image slices range from 1mm to 7mm and the voxel spacing in plane is approximately 0.4×0.4mm. The image size is 512×512. Table 4.1 displays the demographic information of the patients.

The training and validation datasets consist of 781 2D image slices randomly chosen from 101 subjects. 500 of these images were used for training and 281 for validation. There is no patient overlap between training and validation images. A separate test set containing 32 subjects was used. The training, validation, and testing datasets were manually annotated by a human expert. The CSF was segmented into three categories: 1) CSF in the ventricles, 2) CSF in the cerebral cortical sulci, fissures, arachnoid cysts, and 3) other CSF spaces, namely: basal and brainstem cisterns, cerebellar sulci, infratentorial arachnoid cysts. For these image slices, a threshold was chosen to obtain a coarse segmentation on the whole CSF and then the expert edited them using the MRICron software. The suprasellar cistern was bisected, such that CSF anterior to a line joining the bilateral anterior most parts of the cerebral peduncles/midbrain was classified within the cerebral compartment (reflecting atrophy of medial temporal and orbitofrontal cortices, and including Sylvian cisterns); while CSF

posterior to this line (including interpeduncular, crural and ambient cisterns) was classified within the third "cisternal" compartment.

Table 4.1: Demographics of patients in the CSF segmentation experiment.

Age (years)	mean±std	71 ± 14
	range	28–94
Gender	male %	52.63
National Institutes of Health stroke score (NIHSS)	mean±std	10 ± 6.03
	range	1–27

Preprocessing and augmentation: In this work, we do not perform resampling on the CT images. This is because the thickness of the clinical CT images is large (up to 7mm) and resampling the images can introduce inaccuracies and interpolation artefacts. In terms of the image intensity normalization, we employed the similar strategy as described in [24]. We normalized CT images on a per slice basis. This means for each slice, background (i.e. air, bone) was excluded and the remaining intensities (i.e. the HU) were normalized to zero mean and unit deviation. Since we have limited number of image slices for training and one slice is too large for the CNNs, we randomly cropped 128×128 patches from the slice to construct the training set. In this way, the training set contains sufficient number of patches. As our CNNs are fully convolutional, in the testing stage, the input can be the entire image slice.

Baseline: We use the FCN, the U-Net, and the Res-U-Net as baselines. The baseline networks are compared to the DRINet with various growth rates. The results are displayed in Table 4.2.

The FCN and the U-Net perform similarly well in terms of Dice. The results suggest that segmenting the CSF in ventricles is relatively easy while segmenting CSF around brainstem is challenging. As depicted in Figure 4.6, the CSF around brainstem is likely to be misclassified. In addition, the skip connections in the U-Net do not improve the segmentation results in this case.

Changing the U-Net architecture into the Res-U-Net architecture makes the network deeper and reduces the number of training parameters. According to [10], this change should only marginally influence on the results. However, the Dice score of the CSF around brainstem decreases under the Res-U-Net architecture. This result indicates that reducing parameters is problematic although the

Table 4.2: Performance comparison among the baseline CNNs and the DRINet with different growth rates. The numbers under the DRINet indicate the growth rates in each dense connection block.

		Dice (%)				SE (%)			PR (%)			# params
		Ventricles	Cortex	Brainstem	Total	Ventricles	Cortex	Brainstem	Ventricles	Cortex	Brainstem	
FCN	val	83.29	76.71	80.74	84.16	90.17	80.06	79.48	93.47	85.13	83.19	2.71M
	test	92.89	89.01	85.25	90.91	92.86	88.50	86.73	94.76	91.18	84.52	
U-Net [6]	val	82.67	76.10	80.45	84.65	90.07	**83.72**	78.50	93.24	82.60	83.28	2.91M
	test	92.45	89.18	85.20	91.03	92.18	**91.70**	85.31	94.44	88.22	85.73	
Res-U-Net	val	81.66	73.99	76.34	84.15	89.72	79.48	75.84	92.84	85.50	81.67	0.96M
	test	91.64	88.73	82.94	90.76	91.54	87.67	82.43	93.81	91.39	84.34	
DRINet 12,12,12,12	val	84.98	76.87	86.72	82.96	87.24	75.47	76.99	**95.87**	**89.49**	**88.71**	0.85M
	test	92.13	87.77	86.08	89.37	88.76	82.78	82.99	**97.52**	**95.75**	90.29	
DRINet 24,24,24,24	val	85.08	**80.70**	90.87	84.44	91.32	79.67	82.57	93.21	87.12	85.58	2.80M
	test	93.84	89.97	88.40	91.27	**94.78**	88.34	**89.55**	94.27	93.23	87.91	
DRINet 36,36,36,36	val	85.00	80.19	90.08	84.67	89.97	81.73	81.18	94.30	85.57	86.71	5.85M
	test	93.70	90.33	88.48	91.52	92.80	90.23	88.22	96.20	91.93	89.45	
DRINet 48,48,48,48	val	**87.39**	80.00	**91.08**	**84.89**	91.06	82.36	82.18	93.59	85.29	86.74	10.03M
	test	**94.28**	**90.64**	**88.96**	**91.85**	94.19	91.00	89.39	95.55	91.74	89.24	
DRINet 64,64,64,64	val	86.97	79.95	90.58	84.62	90.63	80.51	81.15	93.96	86.64	88.33	17.33M
	test	94.15	90.20	**88.96**	91.53	94.27	88.78	87.43	95.37	93.37	**91.28**	
DRINet 12,24,36,48	val	85.74	79.38	87.92	84.55	90.88	81.81	82.21	93.50	85.40	85.21	4.11M
	test	93.87	90.26	88.15	91.50	93.95	90.32	88.91	95.38	91.77	88.15	
DRINet 24,36,48,64	val	86.98	79.63	90.84	84.69	**93.90**	**85.75**	**87.32**	90.74	81.58	81.30	8.03M
	test	94.27	90.16	88.82	91.51	94.19	89.53	87.83	95.68	92.45	90.53	
DRINet 36,48,64,80	val	86.45	80.08	89.68	84.72	89.86	80.96	82.10	94.58	86.43	87.22	13.70M
	test	93.76	90.27	88.82	91.46	92.44	89.38	88.59	96.64	92.79	89.76	

network uses the residual connections.

The growth rate is the key hyper-parameter in the DRINet because it controls the network parameter space and performance. Changing the growth rate allows to compare the performance between baseline networks and the DRINets with a similar number of parameters. Table 4.2 shows the results evaluating the effects of growth rate. The DRINet with a growth rate of 12 has a similar number of parameters as the Res-U-Net. This DRINet segments the CSF around brainstem significantly better than the Res-U-Net. The DRINet with a growth rate 24 is comparable to the FCN and the U-Net in terms of the size of parameter space. It performs better than the FCN and the U-Net in terms of the CSF in ventricles and around brainstem. If the growth rate increases to 48, the DRINet performs best in all three parts of the CSF segmentation, as well as the whole CSF segmentation. When the growth rate becomes very large (e.g. 64), the DRINet is likely to overfit and the performance decreases. In the following experiments, a growth rate of 48 is used.

Huang et al. [130] noted that a larger growth rate in the higher layers is beneficial for the performance of network. In our experiments, we evaluate this strategy using growth rates like 12, 24, 36, 48 in each dense connection block. Comparing DRINets using identical growth rate and increasing growth rates, which have similar number of parameters, the DRINets using increasing growth rates do not perform significantly better in any part of CSF segmentations.

Run time: Pre-processing was performed on a desktop PC with an Core i7-3770 processor and 32GB RAM. CNNs were trained and tested on an NVIDIA TITAN XP GPU processor except for the DRINets with large growth rates (e.g. 48, 64), which were trained on two GPUs to keep the batch size sufficiently large. On average it took 44.46s for the DRINet to segment the CSF in one image. The training time of the DRINet with the best performance was 21.37 hours. In contrast, the U-Net is faster with 11.44 hours for training and 23.56s per image for testing. Although the DRINet is slower, its run time is acceptable.

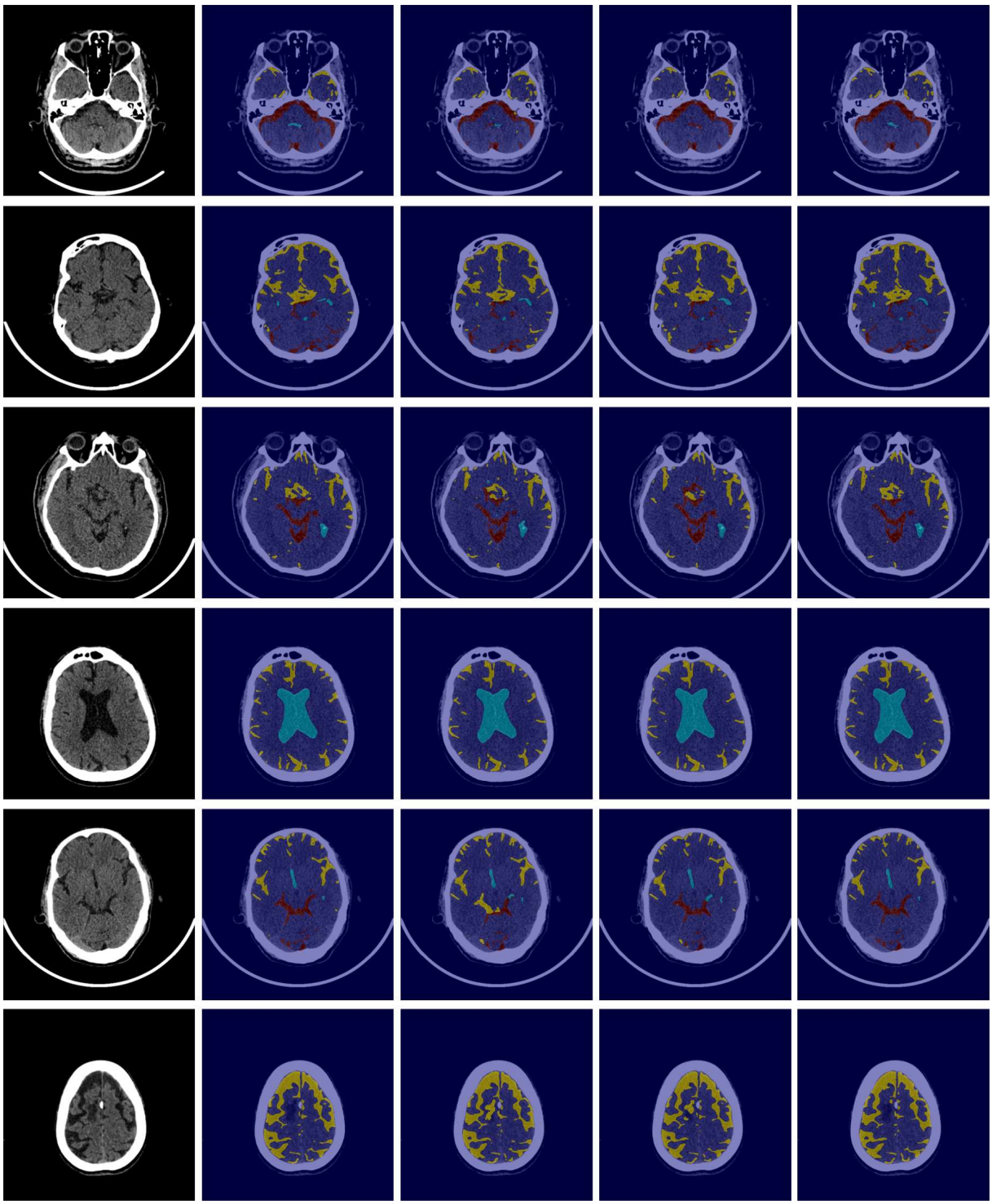


Figure 4.6: The visual examples of multi-class CSF segmentations. The first column displays the original images. The second column shows the manual references. The following columns demonstrate the segmentations of the U-Net, the Res-U-Net, and the DRINet.

4.4.2 Multi-organ Segmentation

Overview: Segmenting abdominal organs is important for clinical diagnosis and surgery planning [147]. For instance, focal lesions on the liver can be detected with the segmentation results [148]. The

kidneys' condition can be measured based on their sizes, which are derived from the segmentations. In addition, diagnosing the dilated pancreatic ducts or inflamed pancreatic tissues requires the pancreas segmentation [149]. There are two major challenges in the multi-organ segmentation problem: 1) these organs have various shapes and sizes; 2) they are mixed together and the borders of organs are weak.

Abdominal organ segmentation is a popular topic for which many solutions have been proposed. Many methods were based on statistical shape models [150] or multi-atlas segmentation [150–154]. Using recent deep learning approaches the segmentation accuracy has significantly improved, particularly for smaller organs (e.g. pancreas). Furthermore, deep learning approaches are much faster than conventional methods [155–157].

Dataset: 3D abdominal CT scans were used in this experiment to evaluate the performance of the DRINet. These images were acquired at Nagoya University hospital using a Toshiba Aquilion 64 scanner and obtained under typical clinical protocols. The image resolution is 512×512 voxels in plane and there are between 238 and 1061 slices per patient depending on the field of view and the slice thickness. The voxel size ranges from 0.55 to 0.82mm and the slice thickness ranges between 0.4 and 0.8mm. All patients were scanned for the purpose of laparoscopic resection of the stomach and gallbladder glands or colon. Three human experts manually segmented the pancreas, kidneys, liver, and spleen on all the images, which was based on the interactive region growing. The demographics of the patients is listed in Table 4.3.

Table 4.3: Patients involved in the multi-organ segmentation experiment.

# subjects	all	150
	training	75
	testing	75
Age (years)	mean$\pm$std	62.80 ± 12.00
	range	26–84
Gender	male %	76

Pre-processing and augmentation were carried out in similar manner to those for CSF segmentation. The only difference is that in the CSF segmentation, the image intensity normalization is performed per slice while in this multi-organ segmentation task, the image intensity is normalized per volume.

We used the same the experimental settings and CNN configurations as in the previous experiments, so no parameters tuning is performed in this experiment. The purpose is to validate the flexibility of the DRINet so this experiment is still based on 2D and the 128×128 image patches were randomly cropped to develop the training set. Therefore, we only split the whole dataset into a training set (75 subjects) and a separate testing set (75 subjects).

Baseline: Again, the U-Net and the Res-U-Net are used as baselines. Table 4.4 displays the segmentation results. The performance of the U-Net and the Res-U-Net is comparable. The Res-U-Net provides better PR but worse SE than the U-Net in segmenting the pancreas and kidneys. As mentioned above, the pancreas is the most challenging organ to segment because of its thin and various structure. The strength of the proposed DRINet is demonstrated by the fact that it is able to segment the challenging organs significantly better than the baseline CNNs approaches.

Comparison with existing methods: We compare the DRINet with existing methods evaluated on the same dataset. [152] and [153] proposed methods based on conventional machine learning approaches. According to the results (displayed in Table 4.5) they have achieved fairly good segmentations in terms of kidneys, liver, and spleen. The method proposed by Tong et al. [153] is much faster than the one proposed by Wolz et al. [152]. The 3D FCN proposed by Roth et al. [156] is the state-of-the-art method based on deep CNNs. It is clear that the 3D FCN achieves significantly better results in the pancreas segmentation. Furthermore the inference time is significantly reduced. However, in terms of the other organs, namely the kidneys, liver, and spleen, the 3D FCN did not offer significant improvements.

The DRINet outperforms the 3D FCN achieving the state-of-the-art based on this dataset. Specifically, it improves the pancreas segmentation further from the 3D FCN. In addition, the DRINet promotes the segmentation on other organs as well. Note that the DRINet is only based on 2D image slices without using 3D contextual information. Therefore, this experiments verifies the DRINet is powerful and robust in the multi-organ segmentation problem.

In addition, we compared the DRINet with the state-of-the-art CNNs, including the Dense V-Net [158] for abdominal multi-organ segmentation on CT images on other datasets. Although the comparison is not completely fair, the DRINet results in the best segmentation results.

Table 4.4: Performance comparison among the U-Net, the Res-U-Net and the DRINet. The DRINet outperformed the baseline CNNs, particularly in terms of the pancreas.

	Dice (%)				SE (%)				PR (%)			
	Pancreas	Kidneys	Liver	Spleen	Pancreas	Kidneys	Liver	Spleen	Pancreas	Kidneys	Liver	Spleen
U-Net [6]	80.09	95.80	94.70	94.72	74.89	**95.86**	92.79	93.13	87.98	95.85	**96.65**	95.98
Res-U-Net	79.09	95.41	96.20	94.71	72.41	93.72	96.15	92.92	**89.49**	**97.28**	96.26	95.94
DRINet	**83.42**	**95.96**	**96.57**	**95.64**	**80.29**	95.84	**96.69**	**95.63**	87.95	96.20	96.47	**96.13**

Table 4.5: Performance comparison among different algorithms. It is clear that the DRINet is superior to the existing methods.

	Dice (%)				Time (h)
	Pancreas	Kidneys	Liver	Spleen	
Wolz et al. [152]	69.60	92.50	94.00	92.00	51
Tong et al. [153]	69.80	93.40	94.90	91.90	0.5
Roth et al. [156]	82.20	–	95.40	92.80	0.07
Gibson et al. [158]	75.00	93.00	95.00	95.00	–
Zhou et al. [159]	62.00	91.00	95.00	92.00	–
Hu et al. [160]	–	95.00	96.00	94.00	–
DRINet	**83.42**	**95.96**	**96.57**	**95.64**	**0.02**

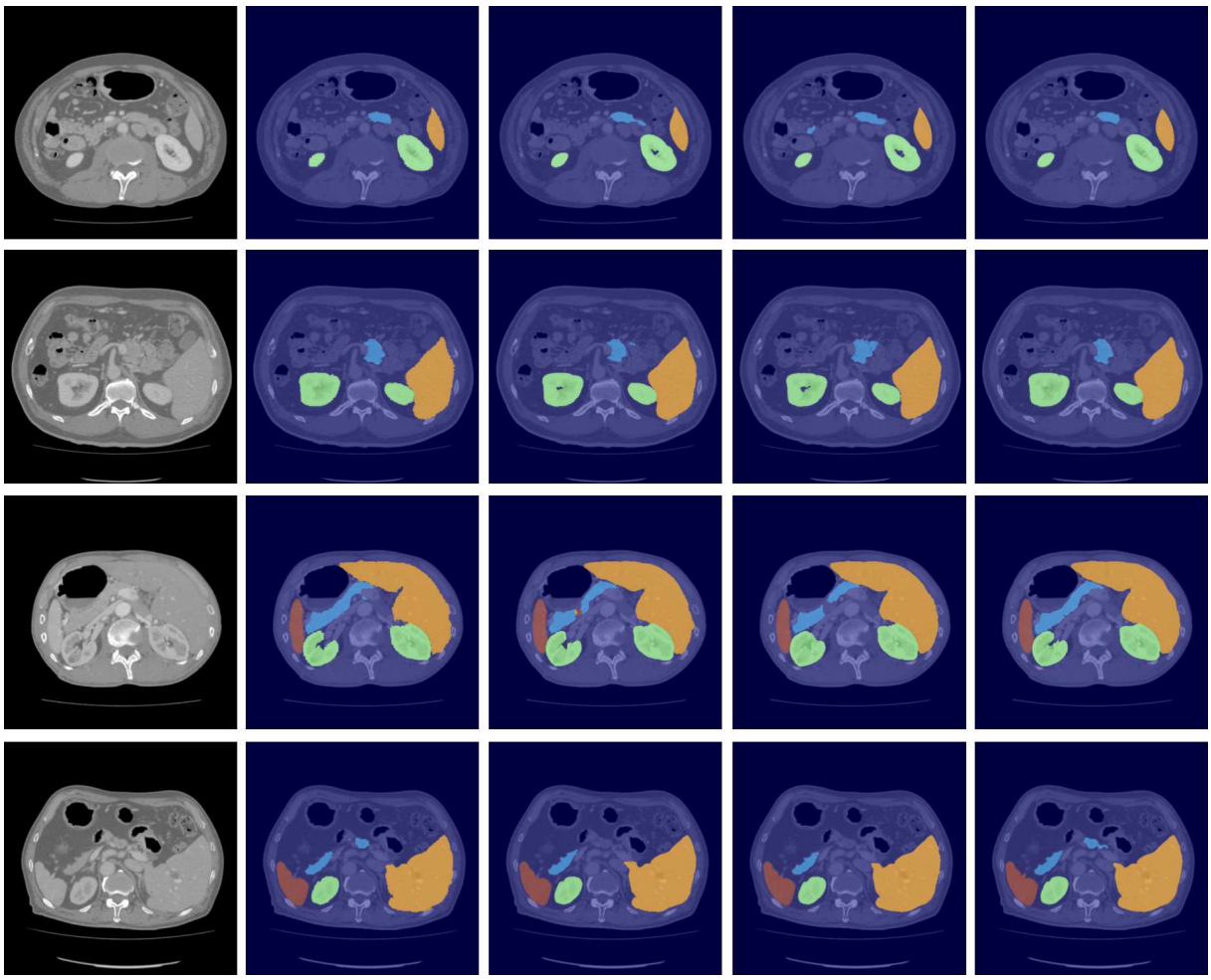


Figure 4.7: The visual examples of abdominal multi-organ segmentations. The first column displays the original images. The second column shows the manual references. The following columns demonstrate the segmentations of the U-Net, the Res-U-Net, and the DRINet.

4.4.3 Brain Tumour Segmentation

Overview: Brain tumours are routinely diagnosed using multi-modal MRI, including native T1-weighted (T1), post-contrast T1-weighted (T1-Gd), T2-weighted (T2), and T2-FLAIR image se-

quences [27]. Quantification of the tumours based on the multi-modal MRI benefits the diagnosis and treatment [161]. Segmenting tumours into necrotic and non-enhancing tumours, the peritumoral edema, and gadolinium enhancing tumours has been a popular research topic [162].

Dataset: We propose to use the training dataset of the BraTS 2017 challenge. There are 285 subjects in total and we randomly select 50 ones for training and the remaining 235 ones for testing. Training on a small number of images is easier to present performance differences between different networks; otherwise different networks are likely to perform equally well. The segmentation is based on 2D patches of size of 64×64. Since the training patch size is smaller compared to that in the previous experiments, all CNNs in this experiments have two downsampling and upsampling process and all the other network configurations are fixed. According to [162], the images have been preprocessed: images were co-registered into the same anatomical template; skulls were stripped; voxels were re-sampled to isotropic resolution ($1mm^3$). We normalise the image intensities into zero mean and unit deviation. No post-processing trick is used in any case. The evaluation is based on the whole tumour region, the tumour core region, and the enhancing tumour core region, instead of individual tumour structures.

Results: On this benchmark dataset, we evaluate the three key components of the DRINet: the dense connection block, the residual Inception block, and the unpooling block. We set the FCN as the baseline CNN and separately add one of the proposed blocks to verify its contribution. We also compare their performance with the U-Net and the DRINet.

Table 4.6 shows the results: In terms of the whole tumour structure, the added blocks do not affect the Dice scores significantly. The dense connection block and the residual Inception block increase the sensitivity and the Hausdorff distances and decrease the specificity, which means they increase the number of FPs. In contrast, the unpooling block decreases the sensitivity and Hausdorff distance and increases the specificity, which means it reduces FPs but introduces FNs. Combining them together results in a trade-off between FNs and FPs. Therefore, the overall performance increases.

In terms of the tumour core and enhanced core, the three blocks increase the Dice scores and specificity while decreasing their sensitivity and Hausdorff distances. This means the overall performance for the segmentation of the tumour core and the enhanced core is improved. However, since their

sizes are fairly small, some FNs occur.

The DRINet with three powerful blocks achieves better segmentation results than the U-Net in terms of the dice scores, the sensitivity, and the Hausdorff distances. Regarding the Res-U-Net, since the parameter space is small, it cannot perform as well as the U-Net in this case. Figure 4.8 shows that the training error of the Res-U-Net is larger than that of the U-Net and the DRINet. Therefore, the dice coefficients given by the Res-U-Net on tumours are the worst among all the CNNs. According to the low sensitivity, the high specificity, and the low Hausdorff distance, it is clear that the segmentation results by the Res-U-Net have many FNs but few FPs.

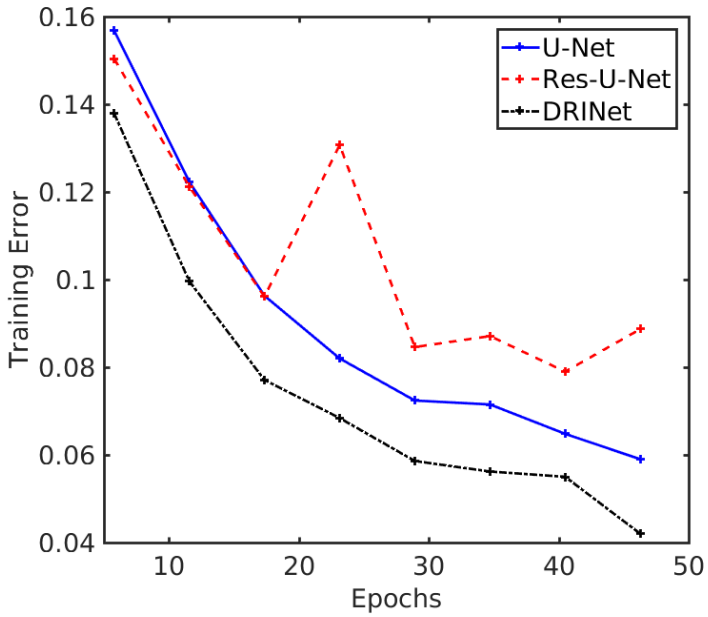


Figure 4.8: The training error comparisons among different CNNs.

4.5 Discussion and Conclusion

In this study, a novel CNN architecture, DRINet, is proposed. The DRINet has three key features, namely the use of dense connection blocks, residual inception blocks, and the unpooling blocks. These blocks deepen and widen the network significantly and the parameter space can be controlled via the growth rate. The gradient propagation is improved due to the dense connections and residual connections. As a result, the performance of the DRINet is significantly improved when compared to the standard U-Net. In addition, the DRINet architecture is highly flexible: Within a block, the

Table 4.6: The segmentation results of different networks. The entries in bold highlight the best comparable results.

Network	Dice (%)			SE (%)			SP (%)			Hausdorff95 (mm)		
	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.
U-Net [6]	81.51	71.30	63.05	81.69	72.51	79.70	99.86	99.92	99.94	42.07	34.44	36.46
Res-U-Net	71.50	67.75	60.06	60.25	66.06	68.27	**99.97**	99.93	**99.97**	**21.98**	**25.00**	27.56
FCN	81.42	70.4	61.49	80.84	**77.12**	**80.76**	99.85	99.80	99.92	42.19	47.24	44.08
FCN+dense	81.09	71.98	63.29	84.90	74.81	78.56	99.80	99.91	99.95	48.34	39.36	36.56
FCN+RI	81.89	72.30	63.25	**85.26**	74.29	78.02	99.82	99.91	99.95	47.38	36.49	33.97
FCN+unpool	81.81	71.43	63.93	78.56	70.53	75.80	99.91	**99.94**	99.96	33.37	28.39	**27.12**
DRINet	**83.47**	**73.21**	**64.98**	84.53	74.93	80.35	99.86	99.92	99.94	36.4	25.59	30.31

convolution/deconvolution layers can be changed adaptively. It is therefore easy to integrate the blocks into other CNN architectures.

In this work, we focus on evaluating the performance of the proposed DRINet and each of its components. The segmentation results of each problem can be improved using some domain knowledge and post-processing. For instance, in the brain CSF segmentation problem, a brain mask could be added. In the abdominal organ segmentation task, 3D contextual information could be included. In the BraTS problem, the CRF model could be used to remove FPs. In both tasks, 3D volumetric convolution can be used to improve the DRINet performance.

Among the three experiments, the multi-class CSF segmentation on CT images is novel. To the best of our knowledge, we are the first to attempt on this problem and the proposed DRINet results in good segmentation. In the future, we plan extend the proposed approach to segment lesions as well as CSF using a single DRINet. This is useful in clinical settings for prognostication after stroke [39] or estimating cerebral hemorrhage risk [40, 41].

In the context of abdominal multi-organ segmentation, the DRINet achieves very good results although the segmentation is based on 2D CT image slices. Our results show that the DRINet improves the segmentation on small and various organs like pancreas as well as big organs like liver. It is of interest to extend its ability to segment more challenging organs such as arteries and veins, which could make the DRINet more useful in clinics.

A limitation of the DRINet approach is that the increase of the growth rate results in many more parameters, which may lead the training more difficult and testing slower. In the future, the research could focus on simplifying the network structure while maintaining its ability.

Chapter 5

Acute Ischemic Lesion Segmentation on DWI

The work in this chapter is based on:

- L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage: Clinical*, vol. 15, pp. 633–643, 2017.

5.1 Introduction

In stroke imaging, the DWI has advantages in diagnosis of acute ischemic lesion in the early stage. The detection and quantification of acute lesions in DWI is important for the diagnosis and treatment of the ischemic stroke. It may allow for accurate estimation of acute lesion volumes. Lesion volume estimation may be important for hyper-acute therapy decision-making, e.g. in determining the ratio of reversible hypo-perfusion to irreversible infarct core [163]. Furthermore, acute lesions can be profiled anatomically in terms of volumes of anatomical-functional regions of interest, by superimposing standard atlas-derived or functional MRI-derived regions [164]. However, manual segmentation of acute ischemic lesions is expensive in terms of time and human expertise. Several automatic and semi-automatic methods have been proposed to assist clinicians to address this problem [165–171]. A common limitation of these models is that they were developed on small datasets which only contain tens of subjects. Since the ischemic lesions can occur anywhere in the brain in various shapes and

sizes (see Figure 5.1) [172], a small dataset makes it difficult to cover the large variation in position, shape, and size. Most of these algorithms are based on multi-modal MRI including T1-weighted, T2-weighted, FLAIR, DWI, and apparent diffusion coefficient (ADC) [168, 173]. Two of them only based on DWI are semi-automatic: The first one is an adaptive thresholding algorithm incorporating a spatial constraint [166]. The fully automatic adaptive thresholding segmentation is likely to fail in cases where there are small lesions and/or lesions in low contract to the normal tissue. Therefore, manual editing was introduced to refine the automatic segmentations. The second one is based on active contours algorithms [167], where before applying the proposed algorithms, image slices with artefacts are manually removed. In addition, human experts mark bounding boxes around the target lesions to initialize the algorithm. To the best of our knowledge, Mah et al. [171] proposed the only fully automated method to segment ischemic damage based on a large DWI dataset. However, their approach was dependent on a reference set of normal brain images and it was only applied to lesions in the occipital lobe.

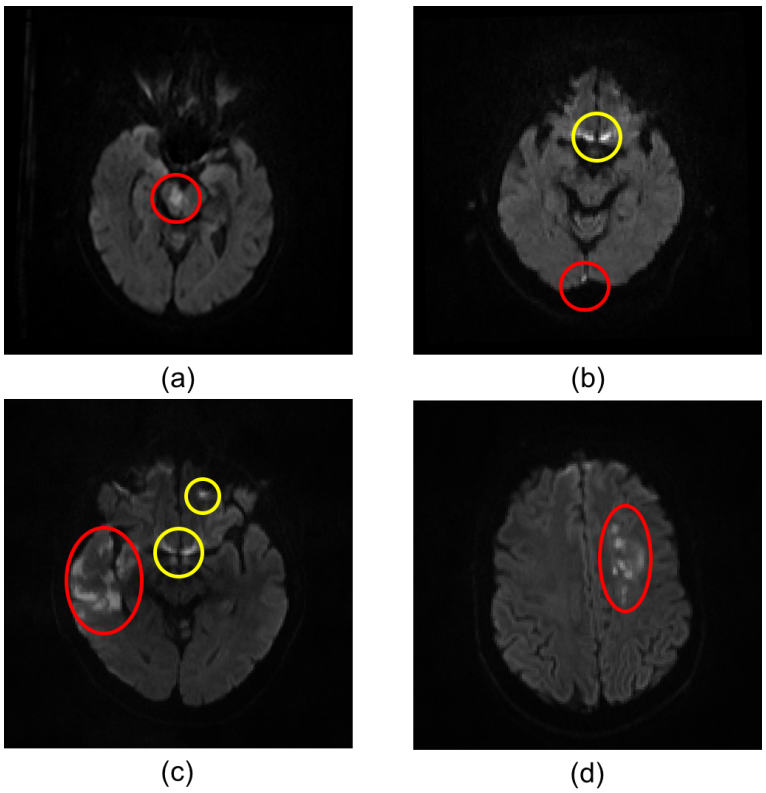


Figure 5.1: Examples of acute ischemic lesions in DWI. The red circles indicate the acute ischemic lesions and the yellow ones show the artefacts.

In clinical practice, semi-automatic methods are still too costly and fully automatic algorithms are

preferred. Although multi-modal images provide rich information about lesions, pre-processing such as resampling and co-registration are required which can lead to inaccuracies. In this study, we propose a fully automatic system (Figure 5.2) to segment acute ischemic lesions in a large DW image dataset based on deep convolutional neural networks (CNNs). Compared to traditional image analysis algorithms, CNNs have major advantages, including end-to-end training and feature learning [174]. Our system consists of two networks, namely the EDD Net and the MUSCLE Net. The EDD Net is an ensemble of two DeconvNets [2] and the MUSCLE Net is the MUti-Scale Convolutional Label Evaluation Net. The input to the proposed system are 2D slices consisting of DWI. The EDD Net firstly outputs a primary segmentation probability map. The binary segmentation obtained by thresholding the probability map contains both lesions and several FPs. The MUSCLE Net re-evaluates all the detections by the EDD Net and excludes many FPs using both the probability map and the original input image.

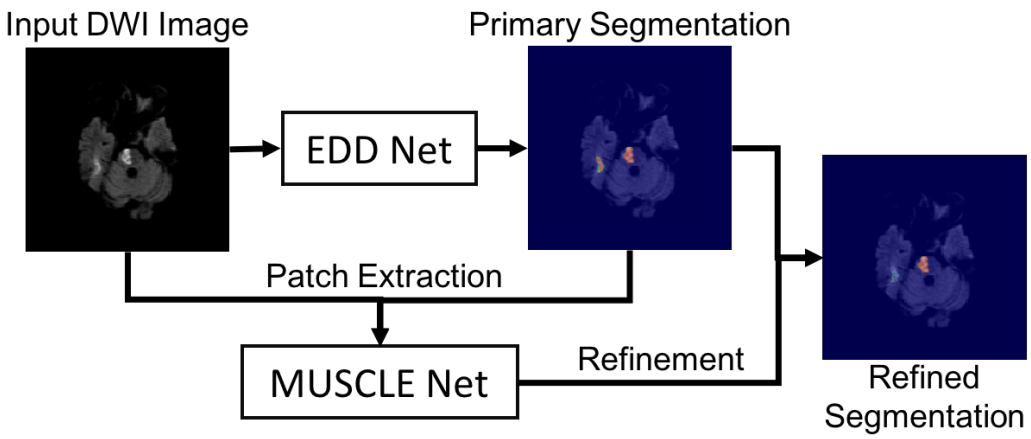


Figure 5.2: The overview of the proposed CNN based system to segment the acute ischemic lesions in DWI. It comprises the EDD Net and the MUSCLE Net. The EDD Net conducts the semantic segmentation on the input DWI. Based on the output of the EDD Net, patches containing small lesions are extracted and they are evaluated by the MUSCLE Net so that many FPs are removed. The refined segmentation is therefore obtained.

The acute ischemic lesion segmentation problem is formulated as a semantic segmentation task. However, the task of semantic segmentation of acute ischemic lesions is different from that of objects in natural images. In natural images, the target objects of interest are dominant in images (e.g. images in the PASCAL VOC [175] dataset) while several acute ischemic lesions can be so small (Figure 5.1 (b)) that they are easy to be overlooked by observers. In addition, it is also difficult to distinguish the boundaries between ischemic lesions and normal tissue (Figure 5.1 (c) and (d)) while objects

in natural images are often characterized by sharp edges to the background. Furthermore, there are many artefacts which have similar appearance to the lesions in DWI (Figure 5.1 (b) and (c)). Air is one of the main resources of these artefacts. They are the major sources of FPs for automated lesion segmentation techniques.

In this study, we propose a novel system to address the ischemic lesion segmentation problem. A key contribution is its ability to handle the lesions of various sizes and shapes while minimizing the number of FPs. Our system achieves the state-of-the-art of the ischemic lesion segmentation performance in DWI while being validated on a large clinical dataset from over 700 patients.

5.2 Related Work

In this section, we review two categories of related work: First, methods that address the BraTS [162] and ischemic stroke lesion segmentation (ISLES) [173] challenges are reviewed. Secondly, we review several CNN-based segmentation approaches that have been recently introduced into medical imaging.

5.2.1 Brain Tumour and Lesion Segmentation

In the BraTS challenges held in 2016, the dataset contains a number of subjects with gliomas and the task is to develop automatic algorithms to segment the whole tumour, the tumour core and the Gd-enhanced tumour core based on multi-modal MR images. In the latest competition [162], over half of the methods were based on DNNs and they achieved top results. For instance, the hyperlocal features (original input image) are used prior to the final segmentation to improve the accuracy [176]. As a pixel-level segmentation problem, there are much more non-tumour pixels than the ones belong to part of the tumours, which means there is a significant label imbalance. To alleviate the imbalance, Lun et al. [177] proposed a re-weighted loss function. Randhawa et al. [178] also modified the cross-entropy loss function so that the segmentations at tumour edges could be improved. Instead of analysing multi-modal MR images in 2D, the DeepMedic approach [179] performs tumour segmentation in

3D while using extended residual connections. In addition to deep learning algorithms, machine learning approaches based on the RF [180–183] also demonstrate good performance using hand-crafted features.

The segmentation of sub-acute ischemic stroke lesion is one of the tasks in ISLES 2015 [173], which attracted many entries. The challenge is to automatically segment sub-acute ischemic stroke lesions based on multi-modal MR images. Compared with the dataset in the BraTS, the dataset used in the ISLES is smaller. Similar to brain tumours, sub-acute ischemic stroke lesions are difficult to segment. In terms of methods proposed, these range from machine learning based methods to deformation based methods. Among the top ranked approaches, DeepMedic [184, 185] was the best, which is a multi-scale 3D CNN with fully connected CRF models achieving a Dice score of 0.59 in testing. The second best performing method used a modified level-set approach embedded with the fuzzy C-means algorithm [186] while the third best method is based on random forests and contextual clustering [187], which is a typical way of segmenting lesions like those in BraTS. They achieved Dice scores of 0.55 and 0.47, respectively. The Dice scores reported by most other attendees ranged from 0.3 to 0.5.

Most of the successful CNN-based methods in both BraTS and ISLES derive a problem specific CNN architecture from generic ones. This is because in these cases, lesions are highly variable in terms of position, size, and shape and artefacts occur frequently. To explore the distinctive lesion features, specific domain knowledge is helpful.

5.2.2 Other CNN-based Approaches to Segmentation

In molecular imaging, a cascaded CNN called deep contour-aware network (DCAN) [188] has been shown to be successful in the gland segmentation task. Prior to the final segmentation, a primary gland object segmentation and a gland contour segmentation are produced separately. The final segmentation is then obtained by fusing the object and contour segmentations. The segmentations are based on multi-level contextual features extracted from the fully convolutional layers. In cell segmentation and tracking scenario, the U-Net approach [6, 68] performs well. In its architecture, the context and

location information of cells are incorporated. In abdominal imaging, multi-level deep convolutional networks have been proposed to segment the pancreas in CT images [189]. This uses a hierarchical coarse-to-fine method studying images from patch level to superpixel/region level. In cardiac imaging, a left ventricle segmentation approach for MR images has been proposed that combines deep CNNs and deformable models [190].

Similar to the deep networks proposed for brain lesion segmentation, generic CNN architectures are often customized for many other medical imaging tasks. However, the U-Net [6] is a generic architecture which can be easily adapted to other cases in medical imaging. More specifically, it is not a task specific method that requires specific prior knowledge (e.g. the input data has to be homogeneous in 3D). Furthermore, since it is a FCN, the input is flexible in terms of sizes and dimensionality.

In addition to the U-Net, the FCN [13] and the DeepLab [14] are another two generic CNNs for segmentations. The FCN [13] is the first CNN which allows end-to-end training for the semantic segmentation problem. It inherits the convolution and pooling layers from contemporary CNNs, including the AlexNet [8], the VGGNet [64], and the GoogLeNet [9], in image classification problems. It adapts them into fully convolutional styles for the semantic segmentation task. The FCN [13] learns features across multiple scales. The DeepLab [14] is a type of improvement to the FCN [13]. In order to gain deep features, the FCN [13] performs many convolutions and poolings which decrease the image resolutions while the DeepLab [14] contributes the atrous convolution and ASPP layers which keep the depth of features without decreasing image resolutions.

5.3 Our Approach

The proposed lesion segmentation framework consists of two modules: The first one is an ensemble of N adapted DeconvNets [2] (EDD Net) (Figure 5.3) and the second one is a MUti-Scale Convolutional Label Evaluation Net (MUSCLE Net) (Figure 5.5). While the EDD Net attempts to achieve optimal lesion segmentation at voxel level, the MUSCLE Net focuses on small lesions that have been detected and aims to remove FPs.

5.3.1 EDD Net

Figure 5.3 shows the architecture of the proposed EDD Net. The input is an image patch, which is fed into N parallel DeconvNets [2] to infer the semantic segmentations respectively. The results from all DeconvNets are then combined. The combination is concatenated with the input image patch. Several convolution layers are added in the end to produce the final output. The CNN is based on 2D to reduce the inaccuracies of image resampling in z-axis.

The basis CNN architecture, i.e. the DeconvNet [2] is selected among several generic CNN architectures for semantic segmentation, including the U-Net [6], the DeepLab [14] and the FCN [13]. The basis network has a stack of convolution and pooling layers in the convolution stage and a stack of corresponding deconvolution and unpooling layers in the deconvolution stage. Within each stack, there are several convolution/deconvolution layers. Between two stacks, there is a pooling/unpooling layer. The number of stacks and the number of layers in each stack define the size of the network. The proposed basis network has three stacks of convolution layers and two pooling layers in the convolution stage, which leads to the best results.

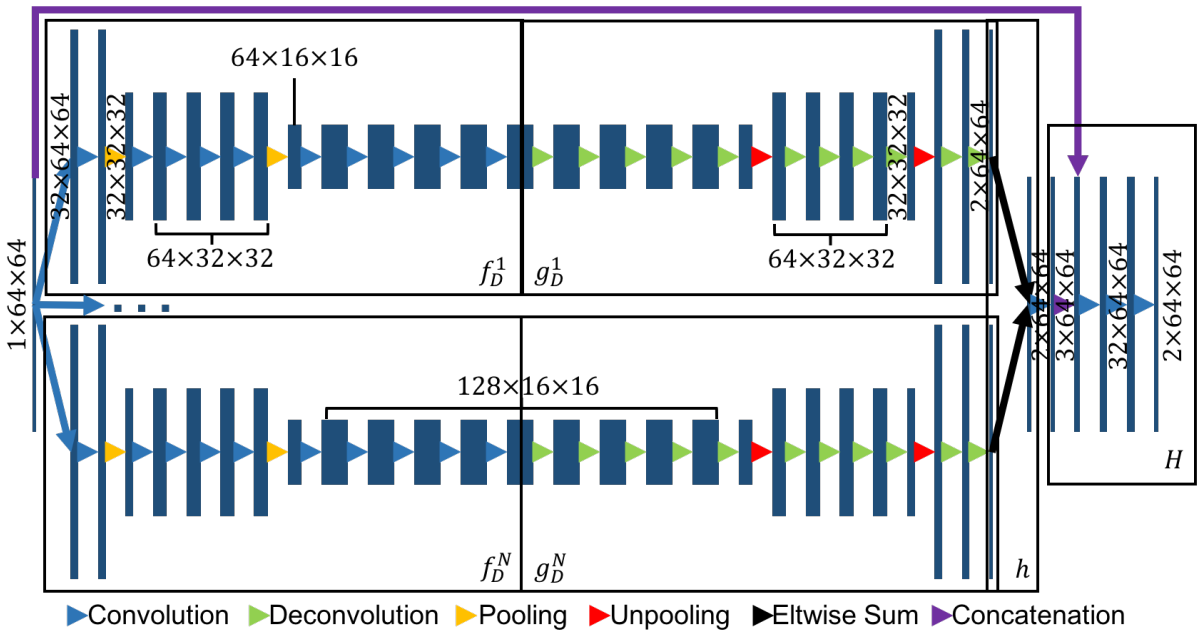


Figure 5.3: The architecture of the proposed EDD Net. The rectangles in different sizes indicate data blobs in different sizes. The height shows the size of each piece of data, e.g. 64×64. The width shows the number of data pieces in each blob, e.g. 1, 32. Arrows in difference colors stand for different operations.

In segmentation problems, contextual information often contributes important knowledge to solve the label assignment. However, the appropriate level of contextual information is often difficult to identify. Excessive amounts of context can hinder the segmentation of lesions and insufficient context makes it difficult to distinguish between lesions and artefacts. If the network grows deep, i.e. has many convolution and pooling layers, it processes a large amount of contextual information. However, with the increasing number of convolution and pooling layers, the input is down-sampled further and further and therefore the resulting feature maps have lower and lower resolutions. In this case, small lesions are gradually eliminated by subsequent down-sampling steps and it can be difficult to reconstruct these. In contrast, if the network is shallow, i.e. using only few convolution and pooling layers, only limited context is used. In this case, lesions and artefacts may have similar feature representations making it difficult for the classifier to distinguish between them.

In our approach, we propose to use image patches instead of image slices as the input. This has three major advantages: Firstly, it modifies the data distribution. For a given image slice, there is a significant imbalance between pixels that represent normal tissues compared to those of lesions since acute ischemic lesions occur locally [33]. The signals representing lesions are as weak as those representing noise and artefacts among the whole data distribution. However, the lesion signals can be apparent among the data distribution based on image patches. Secondly, a large number of patches can be extracted from image slices, which is a fundamental requirement for CNN training. In contrast, if the training data is based on image slices, there is only limited number of candidates available. Finally, as image patches are smaller than image slices, the batch size in training can be larger, which makes the training more efficient [59].

We propose to adopt the DeconvNet [2] as the basis network of the EDD Net. In addition to convolution and pooling layers, the DeconvNet [2] has corresponding deconvolution and unpooling layers to create the segmentation probability map from the coarse feature maps. For the input image patch $\mathbf{x}$, assume $\tilde{\mathbf{x}}$ is the feature maps obtained from the convolution and pooling operations. $f(\cdot)$ and $g(\cdot)$ are the convolution and deconvolution functions which jointly produce the segmentation map $\mathbf{y}$, i.e.

$$\tilde{\mathbf{x}} = f(\mathbf{x}), \mathbf{y} = g(\tilde{\mathbf{x}}).$$

In different architectures, the $f(\cdot)$ functions are similar, which is the composition of several convolutions and poolings, while different strategies are usually used in $g(\cdot)$.

In the DeepLab approach [14], the $g(\cdot)$ function is a bilinear interpolation function upsampling the coarse feature map into the segmentation map directly. In the FCN approach [13], the $g(\cdot)$ not only bilinearly upsamples the feature map but also fuses it with the feature maps obtained at higher resolutions as these contain more image details. Therefore, more small lesions are detected. However, they are difficult to distinguish from artefacts.

In the U-Net [6], the $g(\cdot)$ is modelled in a more sophisticated and powerful fashion. Here, the final segmentation is constructed step by step. In each step, the feature map is upsampled to a higher resolution first, which corresponds to a pooling layer before. The upsampled feature maps are then concatenated with the feature maps before the corresponding pooling layer. Afterwards, a few layers of convolutions are performed on the concatenation. As a result, the segmentation obtained from the U-Net [6] has less FPs than that from the FCN [13] since these convolutions detect and eliminate several FPs.

In the DeconvNet approach [2], there are additional pooling masks $\mathbf{m}$ (Figure 5.4) output from pooling layers who record the locations of the maximal activations. Thus, the specific functions in the DeconvNet [2] can be written as:

$$\tilde{\mathbf{x}}, \mathbf{m} = f_D(\mathbf{x}), \mathbf{y} = g_D(\tilde{\mathbf{x}}, \mathbf{m}).$$

The $g_D(\cdot)$ function represents the deconvolution and unpooling operations. The pooling masks $\mathbf{m}$ are used for upsampling so that the semantic output can be better constructed. Similar to the U-Net [6], the DeconvNet [2] employs a number of deconvolution layers to construct the output step by step, which results in accurate segmentations. In contrast, the U-Net [6] uses feature maps before pooling layers to assist recovering image details, however, this can introduce artefacts and noise. Instead, the pooling masks used in the DeconvNet approach [2] exclude the artefacts and noise.

We propose to combine N DeconvNets [2] to produce an ensemble of classifiers in order to further

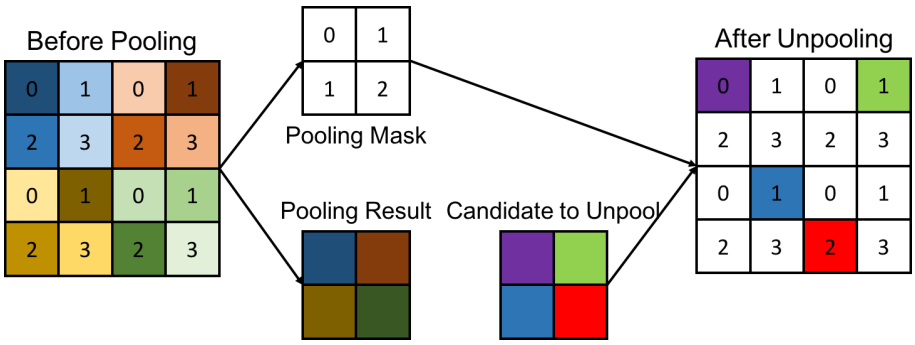


Figure 5.4: The max pooling and unpooling strategy demonstrated in the DeconvNet approach [2]. In the pooling stage, the position of the maximum activation is recorded within each filter window by a mask. In the unpooling stage, the entries are placed in the unpooled map according to the mask.

enhance the results. Let $h(\cdot)$ be the ensemble function fusing the N networks together, i.e.

$$h(\mathbf{x}) = g_D^1(f_D^1(\mathbf{x})) \oplus g_D^2(f_D^2(\mathbf{x})) \oplus \cdots \oplus g_D^N(f_D^N(\mathbf{x})). \tag{5.1}$$

Since the N DeconvNets [2] are initialized differently, they converge at different optima but all of them are able to produce accurate lesion segmentations. An ensemble of all CNNs therefore benefits for performance improvement because of their accuracy and diversity [101]. Furthermore, inspired by the U-Net [6] we propose additional convolution layers at the end of the naive ensemble to refine the segmentation. There are many convolutions and deconvolutions between the original input image and the semantic segmentation. The network may eliminate some details in the input image during the feed-forward pass. We propose to concatenate the input image and the segmentation probability map as well as to add a few convolution layers so that the segmentation can be refined according to the original image. The refinement yields marginal increase of performance. Therefore, the function that the proposed EDD Net performs is

$$H(\mathbf{x}) = r(h(\mathbf{x}), \mathbf{x}) \tag{5.2}$$

Here $r(\cdot, \cdot)$ performs the concatenation and convolutions after the naive ensemble. The loss function

of the EDD Net is therefore

$$\ell = \lambda_1\ell_1(H(\mathbf{x}),\mathbf{y}) + \lambda_2\ell_2(h(\mathbf{x}),\mathbf{y}) + \lambda_3\ell_3(g_D^1(f_D^1(\mathbf{x})),\mathbf{y}) \\ + \lambda_4\ell_4(g_D^2(f_D^2(\mathbf{x})),\mathbf{y}) + \cdots + \lambda_{N+2}\ell_{N+2}(g_D^N(f_D^N(\mathbf{x})),\mathbf{y}). \quad (5.3)$$

In the loss function, $\ell_i(i = 1, 2, \ldots, N + 2)$ is the cross-entropy loss function and the λ_i is the corresponding weight. The loss function is optimised via back-propagation as usual.

The EDD Net is a fully convolutional network since both of its subnets are fully convolutional. Therefore, the size of the input image patch is flexible. In practice, we use the image patches to train the network and we test it on the whole image slice.

5.3.2 MUSCLE Net

The EDD Net identifies many acute ischemic lesions correctly. However, it also produces many false positive (FP) clusters (i.e. aggregation of voxels) which have similar appearance with the small lesions. To remove them, we propose a second network, called MUSCLE Net, which evaluates the labels of small lesions detected by the EDD Net in order to differentiate between FPs and TPs.

The architecture of the MUSCLE Net is shown in Figure 5.5. The input is a stack of image patches across three scales extracted from the original DWI as well as the probabilistic output from the EDD Net. The MUSCLE Net aims at evaluating if the candidate is a real lesion or not. Considering the input patches are fairly small, the MUSCLE Net has limited convolutional layers.

The architecture of the MUSCLE Net is based on a mini VGGNet [64]. The MUSCLE Net consists of four convolution layers, one pooling layer, and three fully connected layers. The convolution and pooling layers extract the distinctive features from the input and the fully connected layers act as a classifier.

The input patch set is derived as follows: First, the primary binary lesion segmentation map is obtained by thresholding the probabilistic segmentation map which is the output of the EDD Net. Based on the binary segmentation map, small candidate lesions are detected using connected-component

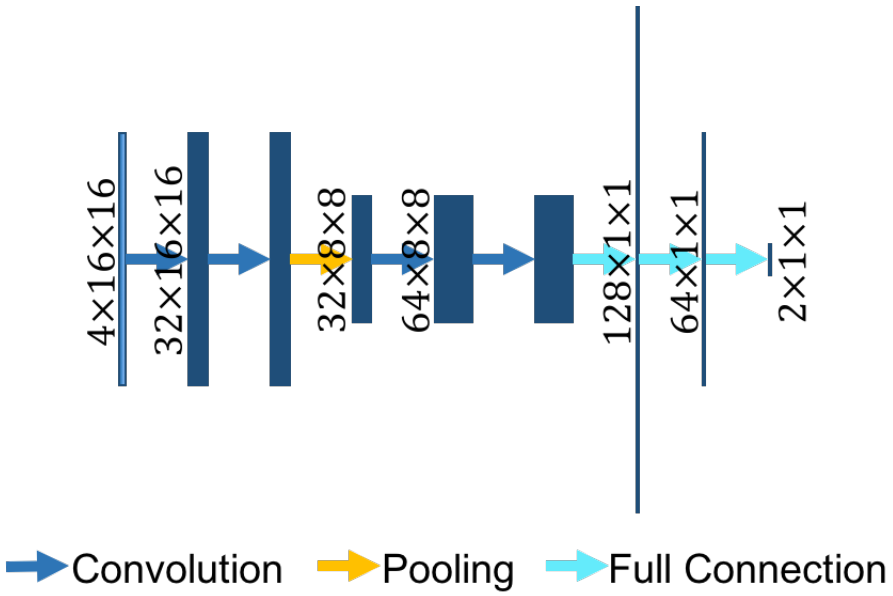


Figure 5.5: The architecture of the MUSCLE Net. The rectangles stand for the data blobs. Their heights represent the sizes of data pieces, e.g. 16×16. Their widths show the number of data pieces in the blobs, e.g. 4, 32. In the fully connected layers, the lengths of strings demonstrate the number of elements in the layers. Arrows in different colors show different operations.

analysis. Original image patches at multiple scales are extracted around them, as well as the corresponding probabilistic segmentation as computed by the EDD Net. This procedure is described in Figure 5.6. The real lesions (TPs) are labelled as positive instances while the FPs are labelled as negative ones.

The MUSCLE Net outputs results at instance level rather than pixel level, which are the probabilities of the candidates being lesions. They are then fused with the pixel level probabilities given by the EDD Net by multiplication. The fused probabilities are re-normalized afterwards. The final semantic segmentation result is therefore achieved. The loss function used here is the cross-entropy function and it is optimised using the BP algorithm.

5.3.3 Evaluation Methods

We propose a number of criteria to evaluate our method. First, the Dice coefficient is used to compare the agreement with manual segmentation. It measures the overlap between the candidate segmentation

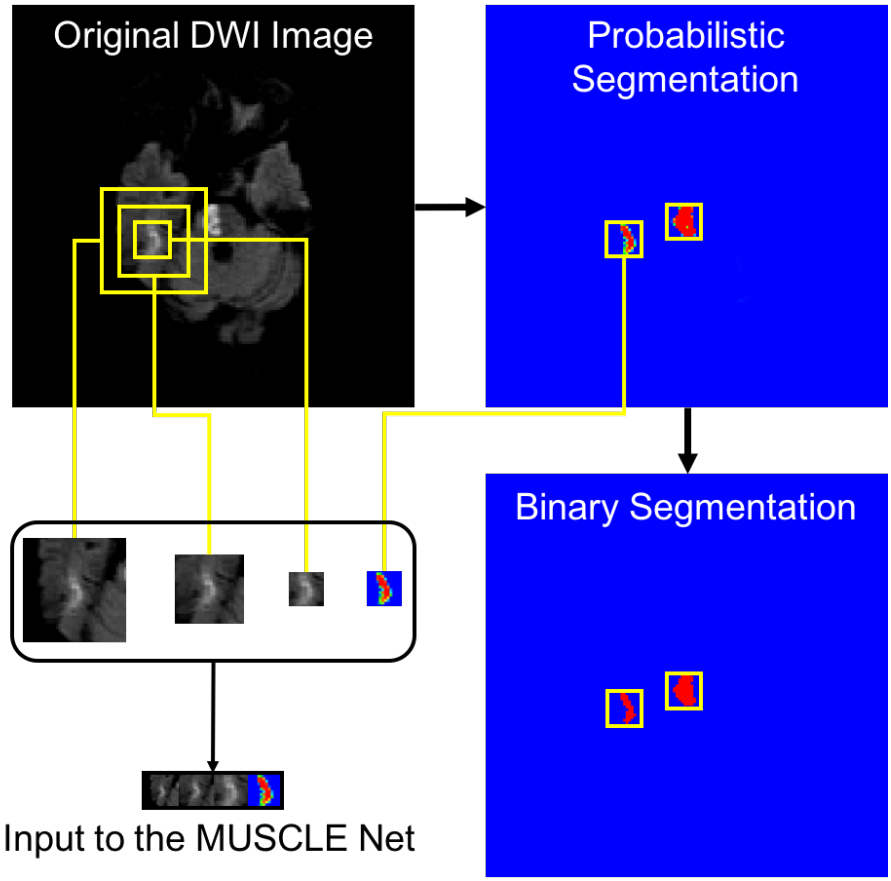


Figure 5.6: The derivation of the input to the MUSCLE Net. The probabilistic segmentation is obtained from the EDD Net. The binary segmentation is obtained by thresholding the probabilistic segmentation. Candidate small blobs are detected in the binary segmentation. The corresponding patches are extracted in the original DWI across multiple scales and the probabilistic segmentation map. They are then resized and concatenated resulting in the input to the MUSCLE Net.

X and the reference segmentation Y and is defined as

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}.$$

$|\cdot|$ denotes the number of pixels in the set. However, the Dice similarity measurement based on overlaps is not robust in all cases: For example, an error of one pixel may not affect the Dice coefficient significantly if the ground truth contains hundreds of pixels; however it makes a significant difference where the ground truth is small and only contains a few pixels. Therefore, the average number (m#) and the average pixel-size (mS) of the FPs and FNs are introduced as additional metrics. Our goal is to decrease the number and size of FPs and FNs. In addition, we define the DR as

$$DR = \frac{N_{TP}}{N}$$

where the N denotes the number of all subjects and the N_{TP} denotes the number of subjects with any true positive (TP) lesion detections. Since the FP may mislead clinicians, the DR is expected to be as high as possible.

5.3.4 Implementation Details

The CNNs in this chapter are implemented using the Caffe framework [191]. The optimization during training is achieved using the standard stochastic gradient descent algorithm. The learning rate is fixed as 0.05. The momentum and the weight decay is set to 0.9 and 0, respectively. The weights in networks are initialized using the Xavier algorithm [192]. The filter size of the convolution and deconvolution layers are 3×3 and the stride is 1. The batch normalization technique [132] is used. We have limited computation resources and therefore set $N = 2$. In the Equation 5.3, we set $\lambda_i = 1, i = 1, 2, \ldots, N + 2$.

5.4 Data

5.4.1 Dataset and Preprocessing

In this study, DWI scans from 741 acute stroke patients were collected from local hospitals. All clinical images were collected from a retrospective database and anonymized prior to use by researchers. Ethical approval was granted by ICL Joint Regulatory Office. The scans were obtained from three different scanners (Siemens) with the following acquisition parameters: field strength: 1.5–3T; slice thickness: 5–6mm; slice spacing: 1.0–1.5mm; matrix size: $(19 - 23) \times (128 \times 128)$ or (192×192); field of view: 230×230 or 267×267; echo time 90-93ms; repetition time 3200–4600ms; flip angle 90°; phase encoding steps: 95–145. Patients information can be found in Table 5.1. In all images the acute ischemic lesions were annotated by experienced experts. We use 380 of them to train and validate our CNNs and the remaining 361 ones are used for testing only. Among the developing images, 274 of them are used for training and 106 ones consist of the validation set.

Table 5.1: Patients information in statistics.

Age (years)	mean: 68.01, std: 14.8, range: 26–93
Gender (male %)	56.28
Interval from acute clinical presentation to MRI (days)	median: 2, std: 1.78, range: 0–9
Admission functional severity (NIHSS)	median: 5, range: 1–30

Since the images were acquired from different scanners under different protocols, several pre-processing steps are performed before experiments. Considering the images are anisotropic in the axial direction (or z-axis) and the resampling is likely to introduce interpolation errors, we will perform analysis of 2D slices instead of 3D volumes. To make sure each pixel in 2D slices has uniform physical pixel size (in mm^2), homogeneous linear resampling is performed in 2D. All images are resampled to uniform pixel size in 2D of $1.6 \times 1.6 mm^2$. Subsequently, the intensity distribution of each image is normalized into that of zero mean and unit variance.

5.4.2 Data Augmentation

Each DWI scan has a limited number of lesions, if the training data is generated at the image slice level or lesion instance level, there is only a small number of images (patches) available. As CNNs have a large number of parameters and it is necessary to generate a large number of images (patches) to train the CNN. For this, data augmentation is implemented in several ways to produce more training data based on the limited number of DWI: First, extracted images (patches) are horizontally flipped and randomly rotated. Second, the patch extraction strategy also represents a way of data augmentation. It is used to reduce the redundant contextual information and balance the number of normal and lesion pixels but it is an effective way of data augmentation. We sample all pixels labelled as part of lesions. For each of these pixels, we extract a patch around it. That pixel is placed in a random position in the patch. As a result, each patch contains pixels belonging to both lesions and tissues/background in general. If the pixel locates in the center of a very large lesion, the patch extracted based on it may contain pixels only belonging to lesions. A pixel cluster of lesions usually have a number of pixels (e.g. 20). That number of patches (i.e. 20) can be generated.

5.5 Experiments and Results

5.5.1 Baseline Architectures

Although the DeconvNet [2] is selected as the basis CNN in the proposed EDD Net, other generic CNN architectures, including the U-Net [6], the DeepLab [14] and the FCN [13], aiming at image segmentation are used as baseline comparison. In this set of experiments, comparisons are among single networks rather than ensembles. The training inputs to all CNNs are patches from the DWI of 64×64 pixel size. This is the best patch size for this task (see Section 5.5.2). Since each architecture has its own characteristics, it is difficult to adapt them so that they have exactly the same size of the receptive field. Fortunately, our results in Secion 5.5.2 show the performance is robust to the size of the receptive field when the image patch size is 64×64. When adapting the candidate CNN architectures into our dataset, we preserve their key features. More specifically, the adapted DeepLab [14] contains atrous convolution and ASPP layers. The adapted FCN [13] is still in the fully convolutional configurations and uses a multi-scale approach. The adapted U-Net [6] has concatenations between related layers. The adapted DeconvNet [2] retains the featured unpooling layer. No post-processing operations such as the CRF are used in any architecture.

The results are displayed in Table 5.2. All CNNs share very high detection rates. The DeconvNet [2] clearly outperforms the other approaches. Since the gap between the U-Net [6] and the DeconvNet [2] is not very significant, we perform paired t-test between them in the testing dataset. The p-value is 1.12×10^{-4}, which indicated that the DeconvNet [2] is superior to the U-Net [6] in this case. As they share similar $f(\cdot)$ functions, the key lies in the $g(\cdot)$ functions. In the $f(\cdot)$ functions, many convolution and pooling operations are performed, which diminishes the activations of lesions in small scales. Basically, all architectures except the DeconvNet [2] employ the bilinear interpolation strategy to upsample the coarse feature maps. This bilinear interpolation makes it difficult to reconstruct the small lesions based on the weak activations. The DeepLab approach [14] produces the output by conducting the bilinear interpolation on the feature maps in the lowest resolution, which introduces many FNs. The FCN approach [13] combines feature maps at multiple resolutions to construct the segmentation map. The feature maps in high resolutions contain signals from small lesions but artefacts and noise

Table 5.2: Performance of the baseline CNN architectures. In each measurement, results on the training, validation, and testing datasets are reported respectively. The DeconvNet [2] is superior to the others in most measurements.

Architecture		DeepLab w/o CRF [14]	FCN [13]	U-Net [6]	DeconvNet [2]
Size of receptive field		44×44	52×52	46×46	44×44
Dice (%)	train	59.85	65.67	71.01	**71.10**
	val	55.10	59.99	**64.13**	61.99
	test	48.08	49.82	52.23	**54.65**
m#FP	train	10.35	11.73	**7.86**	8.32
	val	11.51	13.30	**8.95**	10.08
	test	12.81	16.44	12.85	**11.78**
m#FN	train	4.80	2.96	2.35	**2.19**
	val	4.91	4.00	**3.92**	4.03
	test	5.22	**3.88**	3.99	3.99
mSFP	train	**7.23**	8.40	9.56	8.60
	val	**7.29**	8.66	9.10	8.69
	test	**8.25**	9.92	11.50	10.14
mSFN	train	3.34	2.03	2.17	**1.80**
	val	6.53	5.84	6.20	**5.11**
	test	4.08	3.66	4.17	**3.58**
DR (%)	train	96.73	**99.27**	98.91	98.91
	val	98.11	**99.06**	**99.06**	97.17
	test	92.80	93.63	93.63	**94.18**

as well, which results in a large number of FPs in average. The U-Net [6] is equipped with more powerful operations in its $g(\cdot)$ function so that it performs better than the former two networks. The success of the DeconvNet [2] in this case is due to the recorded pooling masks and the unpooling strategy. They work jointly and are able to preserve the signals from small lesions. Despite the fact that the activations of small lesions are weakened, if they are recorded by the pooling masks, they are likely to be reconstructed in the deconvolution stage. In summary, the pooling mask recording and unpooling strategy works better than bilinear interpolation when there are small lesions.

5.5.2 Patch Size and Receptive Field

The DeconvNet [2] has been validated that it is the best baseline architecture among all candidate CNN architectures. In addition to the CNN architecture, the configuration of the network influences the performance significantly. It is mainly in two aspects which are the size of the input image patches

and that of the network's receptive field. As mentioned before, the size of image patches in the training stage determines the data distribution. The size of the network's receptive field determines the amount of contextual information being considered. They work jointly and experiments in this section aim at discovering how do they affect the CNN's performance.

Single DeconvNets are used in the following experiments. In terms of the input patches, four different sizes are tested. The maximum is the whole image slice. Different sizes of the receptive fields are realized by employing different numbers of convolution and pooling layers. For instance, each DeconvNet branch in the EDD Net (Figure 5.3) has the receptive field in 64×64 pixels.

Table 5.3 displays the results of the DeconvNets [2] for different configurations. It is obvious that when the input patches in the training stage are small (32×32) or large (i.e. the full image size 128×128), the CNN can not perform well in the semantic segmentation task since they contain either insufficient or excessive contextual information. Although small patches can help discriminate the lesions from the normal tissue, which reduces the FNs to the minimum, it is difficult for the network to distinguish between artefacts and the real lesions. As a result, there is a large number of FPs introduced. In the other extreme case where the input is the full image slice, small objects including artefacts and lesions are easily eliminated by the numerous convolutions and poolings. Therefore, few FPs are introduced but there are more FNs. In the mean time, many TPs are ignored by the CNN so that the detection rate falls down. Not surprisingly, patches of medium sizes (64×64 and 96×96) are able to achieve the trade-off between the numbers of FPs and FNs and thus the Dice coefficients on the whole increase to reach an optimum.

The DeconvNets [2] are generally robust to the size of the receptive fields in terms of the Dice coefficient when the size of the training input patches is fixed. Particularly when the patch size is extremely small or large, the overall results are stable in terms of Dice coefficient. In these cases, the size difference of the receptive fields is reflected in the number of FPs and FNs. If the patches are in medium sizes, the Dice coefficient shows little fluctuations. For instance, when the training patches are in 64×64 pixel size, the networks perform similarly to those whose receptive fields are in 32×32 and 44×44 pixel sizes. However, the performance slightly improves when the size of the receptive field increases to 64×64 pixels. When the training patches are in 96×96 pixels, the DeconvNet [2] with

the receptive field in 44×44 pixels has a slightly better performance compared to those with larger receptive fields.

According to the results, the configuration providing the best performance is chosen as the basis network of the EDD Net. More precisely, the training patches are in 64×64 pixel-size and the same as the receptive field. In summary, the training patch size affects the networks' performance more than the receptive field. Patches of medium sizes are preferable. Once the size of training patches is fixed, the network is fairly robust to the size of the receptive field.

5.5.3 Ensemble and Refinement

To further improve the performance, the EDD Net is developed based on the DeconvNets [2] under the best configuration. Table 5.4 displays the results in details. First, the two DeconvNets [2] both provide accurate segmentations as before. Note that the Dice coefficient of them in this experiment were 0.56 which is slightly lower than it in Table 5.3. It is the fact that training two networks simultaneously is more difficult than a single one as the number of parameters doubles. Therefore, the loss function is more difficult to optimise. Second, it is obvious that the naive ensemble of the two networks leads to a significant improvement. This is due to a sharp reduction of the FPs, which results from the diversity of the two DeconvNets [2]. As both of them have detected most of the lesions, the diversity indicates FPs given by them are different. Fusing them together should be able to decrease a substantial number of FPs.

Finally, a few convolution layers are added to refine the segmentation provided by the naive ensemble. The naive ensemble of the two DeconvNets [2] is so deep that the input patches are likely to lose details when being fed forward. Inspired by the U-Net approach [6], concatenating the original input and the result given by the naive ensemble and adding a few convolution layers yields a refined segmentation. In summary, the ensemble based on the accuracy and diversity of sub-nets makes a significant improvement to the network performance entirely.

5.5.4 The MUSCLE Net

The EDD Net has advantages to segment the acute ischemic lesions in DWI. However, FPs are difficult to avoid. We validate the trained EDD Net on the validation dataset and report the FPs in Figure 5.7. Approximately 99% FPs are of size 60 pixels or less. According to the Table 5.4, the FPs on the validation dataset are in 8.87 pixels in size on average. Therefore, the MUSCLE Net is only needed to assess candidates within 60 pixels or less in size, which is defined as small objects.

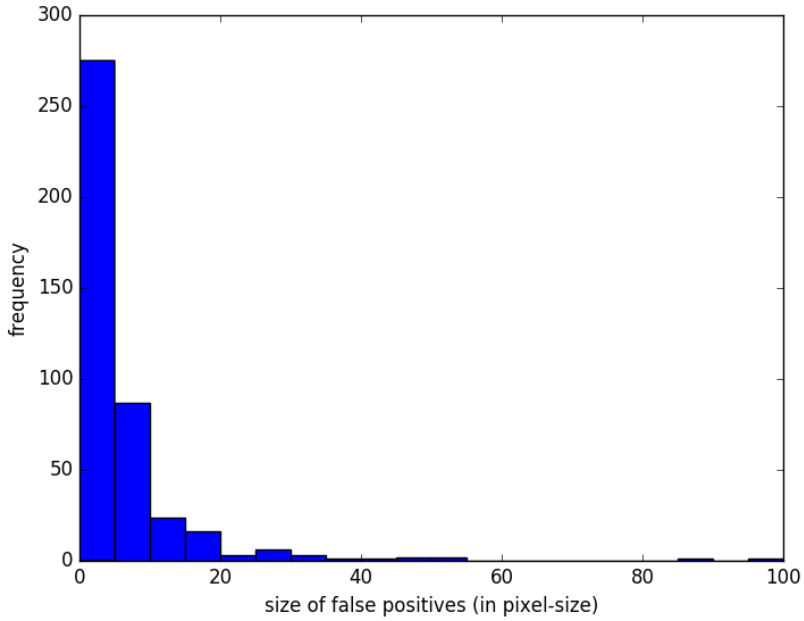


Figure 5.7: The statistics of the FPs on the validation dataset provided by the EDD Net.

Table 5.4 also shows the results of the EDD+MUSCLE Nets. The MUSCLE Net eliminates a large number of FPs without erasing many TPs, which benefits further improvement in performance. According to our observations, the FPs normally appear isolated without overlap with other lesions. Examples are shown in Figure 5.6 and Figure 5.8. This should be one of the major reasons leading to the success of the label evaluation. Although FPs are removed, their mean size grows, which indicates that most FPs within a few pixel-size are eliminated while some slightly larger ones are remaining. The limitation of the MUSCLE Net is that it is not possible to be integrated with the EDD Net to enable the end-to-end training since the training data generation operation is not differentiable. In summary, the MUSCLE Net is powerful to remove FPs without introducing many FNs.

5.5.5 Small and Large Lesions

Apart from the analysis based on the whole testing dataset, it is also interesting to study the performance of our proposed CNNs on datasets with only small or large lesions. First, we compute the mean size of lesions of each subject in our testing dataset and took an average across all subjects. As a result, the mean average size of lesions of the testing subjects is 36.21 pixel-size. Therefore, we regard subjects with average lesions smaller than 37 pixel-size as the ones with small lesions; otherwise with large lesions. Second, the testing dataset is separated into two subsets: one contained subjects with small lesions and the other one consisted of subjects with large lesions. The former subset has 271 subjects and the latter one has 90 subjects. Third, we evaluate our baseline CNN architectures and proposed EDD and MUSCLE Nets based on the two subsets.

Results are displayed in Table 5.5. Not surprisingly, the performance of all CNNs drop down when there are only small lesions. When there are only large lesions, the detection rates were 100%. However, the EDD Net performs significantly better than any of the baseline CNNs. Its mean Dice score is 9% higher than the best baseline CNN. This improvement comes from the significant reduction of the number of FPs as its m#FN, mSFP, and mSFN are similar to the baselines'. In addition, the MUSCLE Net further removes nearly half of the FP artefacts. Importantly, the m#FN of the MUSCLE Net only increases a bit compared to the EDD Net, which indicates that it maintains most of the TP lesions. In terms of the subjects with large lesions, the Dice score achieved by the EDD Net reaches 83%. In this condition, although the MUSCLE Net is still able to remove some small FPs, it can not reflected on the Dice score. The detection rates indicate that when there are large lesions, they can never be ignored by our CNNs. The proposed CNNs may only ignore a few small lesions.

5.5.6 Running Time

The preprocessing computation was run on a desktop PC, which is an HP Elite 8300, with an i7 processor and 16GB RAM. The CNNs were trained and tested on an NVIDIA Tesla K80 GPU processor. We tested the running time of each stage of our proposed pipeline and the results were shown in Table 5.6. In summary, to test a new DWI scan, it costs less than one second, which is very fast.

5.6 Discussion and Conclusion

In this study, we have presented a novel framework based on deep CNNs to segment the acute ischemic lesions in DWI. To the best of our knowledge, it is the first fully automatic method developed for this problem. The algorithm is validated on a large real clinical dataset and achieves the state-of-the-art, which is 0.67 in terms of the Dice coefficient in average. Several visual examples of the segmentation results are shown in Figure 5.8.

Although the combination of EDD+MUSCLE Nets achieves very good results, the proposed approach still has a few limitations: First, semantic segmentation of objects in images across multiple scales remain a challenge that it is not fundamentally solved. Second, the training and testing is not end-to-end, which decreases the system's efficiency. Finally, in the second stage, we only consider the FPs. However, there are still a small number of FNs which must be corrected.

In the future, further improvements could be achieved in several aspects. In particular, more DW images should be collected for training and testing. Our method is capable of automatically generating acute ischemic lesion segmentations. Experts could create the manual annotations based on the automatic segmentations, which will be less expensive in terms of time and effort. In addition, the framework could be adapted so that the end-to-end training is possible. Last but not least, convolutions in our proposed networks could be extended to 3D, which may reduce more FPs. 3D convolutions require the image patches and/or volumes to be isotropic in 3D [184, 185]. However, image slices in our dataset are very thick and simple processes such as resampling cannot provide satisfactory results. Therefore, we consider to employ image super resolution techniques [193] to enhance the images in 3D. Then 3D convolutions can be used in our CNNs.

Table 5.3: Results of the DeconvNet [2] in different configurations. In each measurement, results on the training, validation, and testing datasets are reported respectively. It is clear that the size of training patch size influences on the performance more than the size of network's receptive field.

Size of input patch		32×32		64×64			96×96			128×128		
Side length of receptive field		18	32	32	44	64	44	64	96	64	96	128
Dice (%)	train	47.98	48.56	70.71	71.10	**74.32**	71.88	68.96	67.64	62.36	62.63	60.84
	val	43.54	43.77	**64.26**	61.99	63.65	62.65	58.94	57.59	49.99	52.57	50.78
	test	35.89	36.41	54.75	54.65	**58.01**	54.37	51.98	50.76	46.58	48.37	47.37
m#FP	train	44.32	38.16	9.09	8.32	5.41	8.53	9.69	12.93	1.68	1.82	**0.96**
	val	43.14	38.96	11.04	10.08	7.88	11.26	12.90	16.08	2.75	2.64	**1.63**
	test	51.23	41.07	12.82	11.78	7.92	13.74	13.18	17.39	3.45	3.41	**1.75**
m#FN	train	2.74	2.63	2.62	2.19	2.12	2.35	**1.93**	1.97	5.40	5.33	5.59
	val	**3.17**	3.41	3.97	4.03	4.39	4.09	4.50	4.41	6.37	6.19	6.52
	test	**2.82**	3.31	3.82	3.99	4.25	3.95	4.26	4.14	6.53	6.41	6.83
mSFP	train	9.34	10.42	6.97	8.60	9.30	7.05	8.73	7.37	3.25	5.10	**2.97**
	val	9.73	10.20	6.51	8.69	8.52	7.29	8.37	7.20	4.07	5.66	**3.10**
	test	10.41	11.30	8.05	10.14	10.63	7.79	9.81	8.01	4.81	6.34	**4.40**
mSFN	train	2.17	2.49	2.21	1.80	1.99	1.99	1.64	**1.57**	3.19	3.01	3.33
	val	4.12	**3.53**	6.67	5.11	7.48	6.00	5.46	6.44	8.18	7.94	8.38
	test	**3.02**	3.47	4.05	3.58	3.70	3.77	3.94	3.53	5.54	5.23	6.22
DR (%)	train	**99.27**	98.18	98.54	98.91	98.18	98.91	**99.27**	98.91	97.82	97.82	96.73
	val	**99.06**	**99.06**	**99.06**	97.17	**99.06**	**99.06**	**99.06**	98.11	96.23	95.28	95.28
	test	**95.29**	93.63	93.63	94.18	93.91	94.46	93.07	93.91	90.30	90.58	91.14

Table 5.4: Results of the EDD and the MUSCLE Nets. In each measurement, results on the training, validation, and testing datasets are reported respectively. The ensemble contributes a significant improvement to the whole performance. The MUSCLE Net shows its advantage in removing FPs to boost the performance tremendously again.

		DeconvNet 1	DeconvNet 2	Naive ensemble	EDD Net	EDD+MUSCLE Net
Dice (%)	train	74.42	72.48	79.07	80.41	**88.39**
	val	63.98	61.42	67.60	68.74	**72.81**
	test	56.38	56.18	61.66	62.56	**66.71**
m#FP	train	6.82	9.49	4.20	3.78	**0.64**
	val	9.23	12.27	6.33	5.67	**3.14**
	test	10.18	13.38	6.68	5.89	**3.27**
m#FN	train	1.80	1.59	1.51	1.45	**1.45**
	val	4.08	**3.80**	4.02	4.01	4.16
	test	4.02	**3.66**	3.81	3.82	4.07
mSFP	train	8.39	**6.89**	9.55	9.49	8.81
	val	8.09	**7.33**	9.01	8.87	8.95
	test	9.55	**7.37**	10.31	10.53	12.16
mSFN	train	1.86	**1.40**	1.41	1.42	1.42
	val	**5.58**	5.71	5.65	5.62	6.32
	test	3.81	**3.19**	3.49	3.64	4.16
DR (%)	train	**99.27**	98.55	98.91	**99.27**	**99.27**
	val	99.06	99.06	99.06	99.06	99.06
	test	93.91	**94.46**	94.18	94.18	**94.46**

Table 5.5: Performance comparison among adapted existing CNNs and our proposed CNNs on two subsets of testing dataset. One subset consisted of 271 subjects with small lesions and the other one contained 90 subjects with large lesions. The results showed the EDD Net performed significantly better than existing CNN architectures, particularly on the first subset. The MUSCLE Net further improved it by removing more FPs while maintaining TPs.

		Dice (%)	m#FP	m#FN	mSFP	mSFN	DR (%)
DeepLab w/o CRF [14]	small	39.13	12.84	4.96	**8.16**	3.52	90.41
	large	75.03	12.72	6.00	**8.52**	5.80	100.00
FCN [13]	small	40.73	16.74	3.63	9.81	**3.16**	91.51
	large	77.19	15.56	4.62	10.23	5.16	100.00
U-Net [6]	small	43.50	12.81	3.80	11.75	3.61	91.51
	large	78.52	12.97	**4.56**	10.73	5.87	100.00
DeconvNet [2]	small	46.71	11.38	3.75	10.21	3.21	92.25
	large	78.58	12.98	4.72	9.92	**4.72**	100.00
EDD Net	small	55.91	5.58	**3.58**	10.59	3.17	92.25
	large	82.59	6.82	**4.56**	10.38	5.06	100.00
EDD+MUSCLE Net	small	**61.18**	**2.97**	3.83	12.58	3.68	**92.62**
	large	**83.37**	**4.16**	4.78	10.90	5.58	100.00

Table 5.6: Running time of our proposed pipeline. The unit of time in testing is second and it in training is hour. The numbers in testing are in the form of mean±std while the training time was measured in once.

	Running Time	
	Testing (s)	Training (h)
Preprecessing	0.20 ± 0.10	–
EDD Net	0.63 ± 0.07	26.61
Muscle Net	0.07 ± 0.05	0.11
Total	0.90 ± 0.12	26.72

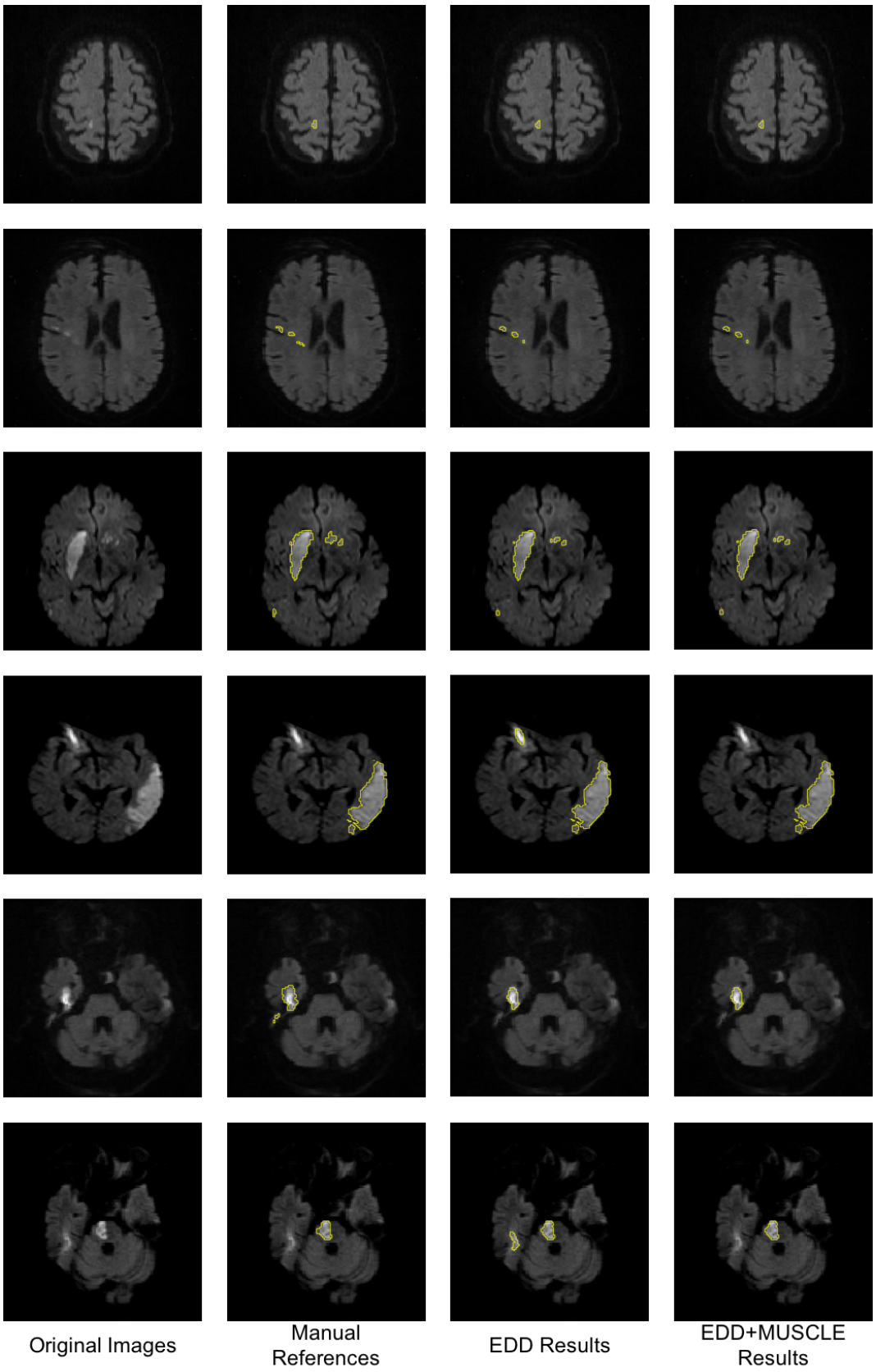


Figure 5.8: The results of the proposed method. The first column shows the original DWI. The second column displays the manual annotations of the acute ischemic lesions. The third column demonstrates the results given by the EDD Net. The last column illustrates the lesion segmentations refined by the MUSCLE Net.

Chapter 6

Self-Supervised Feature Learning for Medical Image Analysis

The work in this chapter is under review on:

- L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised feature learning for medical image analysis," *Medical Image Analysis*, 2018.

6.1 Introduction

Deep convolutional neural networks (CNNs) have achieved great success in computer vision, including image classification [8, 9, 64], object detection [194, 195] and semantic segmentation [13, 14]. In medical image analysis, CNNs have also demonstrated significant improvement when applied to challenging tasks such as disease classification [196, 197] and organ segmentation [6, 68, 185]. Large amounts of training data with manual labels have been crucial in many of these successes. In natural images, crowdsourcing can be used to obtain ground-truth labels for the images [63]. This is based on the fact that the annotation of natural images only requires simple human knowledge, e.g. most humans are able to recognize cars in natural images. However, crowdsourcing has limited applicability in medical imaging because annotation usually requires expert knowledge. This means it is usually

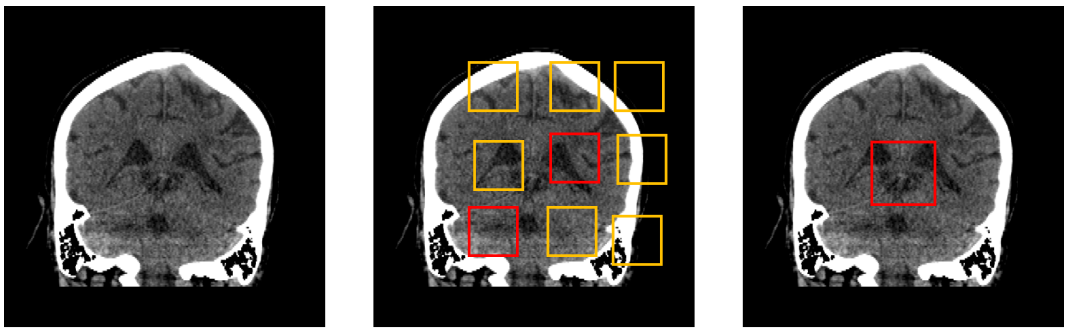
easier to access a large number of unlabelled medical images rather than a large number of annotated images.

Training CNNs only using the small number of labelled images cannot always achieve satisfactory results and does not exploit the potentially large number of unlabelled images that may be available. The most straightforward method to make use of unlabelled data is to train an auto-encoder [3] to initialise the task-specific CNN. However, the loss function used in auto-encoder is the L2 reconstruction loss which leads the auto-encoder to learn features that have limited value for discriminative tasks.

Self-supervised learning is a type of machine learning strategy which has gained more and more popularity in recent years. It aims at supervised feature learning where the supervision tasks are generated from data itself. In this case, a very large number of training instances with supervision is available. Pretraining a CNN based on such self-supervision results in useful weights to initialise the subsequent CNN based on data with limited manual labels. Therefore, self-supervised learning is a good option to explore the unlabelled images to improve the CNN performance in case where only limited labelled data is available.

In this study, we focus on self-supervision for medical images. Two existing self-supervised learning strategies are applicable in our cases, namely, the prediction of the relative positions of image patches [4] (the RP method) and local context prediction [5] (the CP method). Figure 6.1 shows an example of these two methods. In the RP approach, a 3×3 patch grid is selected and the CNN learns the relative position between the central patch and one of its surrounding patches. For instance, a patch containing left cerebellum should locate at the bottom left corner of the patch of right cerebrum. In the CP method, a patch in the centre of image is selected and a CNN learns to predict its context using other image context.

We propose a novel self-supervised learning strategy for medical imaging. Our approach focuses on context restoration as a self-supervision task. Specifically, given an image, two small patches are randomly selected and swapped. Repeating this operation a number of times leads to a new image for which the intensity distribution is preserved but its spatial information is altered. A CNN is then trained to restore the altered image back to its original version. The proposed context restoration strat-



(a) Original image (b) RP method (c) CP method

Figure 6.1: Demonstration of the RP and CP method on a brain CT image. (a) shows the original CT image in the coronal view. (b) shows the patch grid of the RP method and the red rectangles indicate patches of left cerebellum and right cerebrum. (c) shows the selected patch to be predicted.

egy has three advantages: 1) CNNs trained on this task focus on learning meaningful features; 2) CNN weights learned in this task are useful for different types of subsequent tasks including classification, localization, and segmentation; 3) implementation is simple and straightforward. We evaluate our novel self-supervised learning strategy in three different common problems in medical image analysis, namely classification, localization, and segmentation. Our evaluation uses different types of medical images: image classification is performed on 2D fetal ultrasound (US) images; organ localization is tested on abdominal computed tomography (CT) images; and segmentation is performed on brain magnetic resonance (MR) images. In all three tasks, the pretraining based on our context restoration strategy is superior to other self-supervised learning strategies, as well as no self-supervised training.

6.2 Related Work

The key challenge for self-supervised learning is identifying a suitable self supervision task, i.e. generating input and output instance pairs from data. In computer vision, various types of self supervision have been proposed depending on data types and target task, which is summarised in Table 6.1.

For static images, patch relative positions [4, 87], local context [5], and colour [86, 198] have been used in self-supervised learning. In the RP method, it was proposed to predict the relative positions between a central patch and its surrounding patches in a 3×3 patch grid [4]. The idea was that there are intrinsic position relations among divided parts of an object of interest. The RP method has three shortcomings: First, the relative position between two patches could have multiple correct answers,

Table 6.1: Summary of related literature. There are many self-supervision strategies have been proposed for natural images and videos while there is only one strategy relating to medical images.

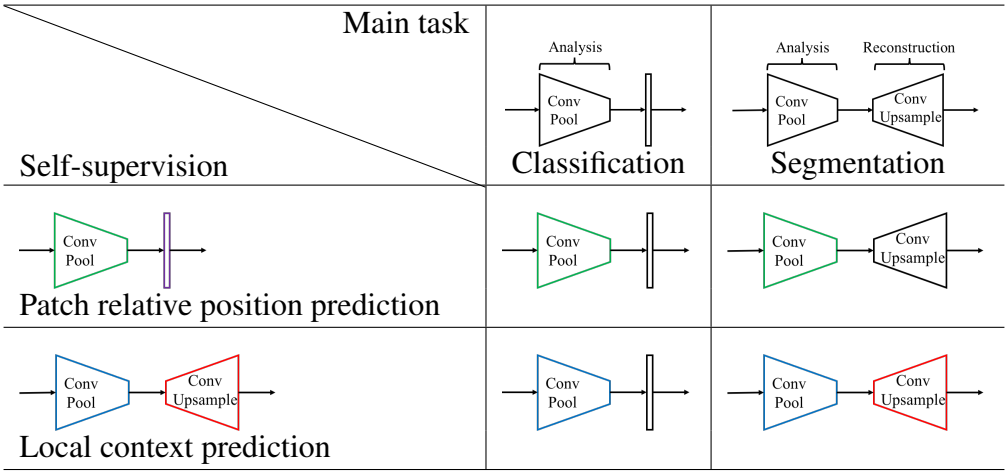
Data Type	Authors	Supervision
RGB images	Doersch et al. [4] Noroozi et al. [87]	patch relative position prediction
	Pathak et al. [5]	local context prediction
	Zhang et al. [86]	colourization
	Zhang et al. [198]	colour-context cross prediction
	Dosovitskiy et al. [199]	exemplar learning
Videos	Mobahi et al. [90] Jayaraman et al. [200]	temporal coherence
	Wang et al. [201]	temporal continuous
	Walker et al. [202] Purushwalkam et al. [203] Sermanet et al. [204]	object motion prediction
	Misra et al. [205] Fernando et al. [206]	temporal order verification
Multi-modal data	Agrawal et al. [88] Jayaraman et al. [207]	ego-motion prediction
	Owens et al. [208] Chung et al. [209]	audio-video matching
MR images	Jamaludin et al. [210]	follow-up scan recognition

e.g. a patch of a car and a patch of a building. Second, it was reported that CNNs could complete the self-supervised learning tasks by learning trivial features, instead of meaningful features that are useful in other discriminative tasks such as classification and segmentation. Specifically, in the RP method, CNNs learns the shared edges or corners of two patches to predict their relative positions. Although techniques were proposed to address this effect, CNNs could still learn trivial features. For instance, it was proposed that patches are randomly jittered so that there is no shared information at edges or corners. However, the CNN may still learn patch positions from some background patterns. Third, the RP method is based on patches, which do not convey information about the global context of images. As a result, the RP method can only provide limited improvements for subsequent tasks requiring global context, such as classification. Later, a more complicated version of patch relative positions was proposed [87], in which all 9 patches are input to CNNs in a random sequence. The CNNs were trained to find the correct sequence of the patches.

In terms of feature learning, learning to predict image context is more straightforward as proposed

by Pathak et al. [5]. They proposed an idea which trains CNNs to learn how to inpaint missing information in images with patchy context removed. For the inpainting, an adversarial loss was proposed in addition to the L2 reconstruction loss while for feature learning only the L2 loss was used. They reported that if the removed patch is always in the centre of an image and in the square shape, CNNs would only focus on the central context. As a result, patches with random shapes and in random locations were removed to improve the feature learning. However, the removal of context changes the image intensity distribution. Thus the resulting images belong to another domain and the learned features may not be useful for images in the original domain. Compared to the RP method, the CP method is more useful for the subsequent tasks. More precisely, the CNN weights learned in the CP method can be used to initialise subsequent CNNs for classification and segmentation; while CNN weights learned in the RP method can initialise subsequent classification CNNs and only the analysis part of the subsequent segmentation CNNs. This is because a CNN predicting the relative positions of patches is a classification model, which does not have layers to reconstruct image-level maps. Table 6.2 compares the RP method and the CP method in terms of subsequent task initialization.

Table 6.2: Comparison between the RP method and the CP method. Weights learned in both of them can initialise the subsequent classification CNN. Weights learned in the RP method can only initialise the analysis part of the subsequent segmentation CNN; while weights learning in the CP method can initialise analysis and reconstruction part of the subsequent segmentation CNN.



Colour is one of the most important features in natural images. It was proposed that learning colours from greyscale images learns features that capture semantic information [86], i.e. CNNs must implicitly perform object recognition in order to colour them appropriately. However, it is generally difficult

to recognize if the weather is sunny or not in greyscale images. Therefore, learning semantics via colours is difficult to cover all aspects of object variance. In subsequent work, Zhang et al. [198] proposed stronger supervision. Specifically, natural images were firstly converted into greyscale space and colour space. Then image representing each space was used to train a siamese CNN to predict the information in the other space. Combining the two outputs reconstructs the original image. This cross-supervision forces the CNNs to learn more meaningful semantics. In medical imaging, most images are in greyscale so that no colour information is available.

In addition, the exemplar learning has been proposed as a self-supervised learning strategy [199]. In exemplar learning, the task is to classify each data instance into a unique class. In this case, heavy augmentation is required to generate training data. Since each data instance is regarded as one class, the exemplar learning method is difficult to apply to large datasets.

Image sequences (or videos) offer rich resources which could be used in self-supervised learning. First, neighbourhood frames should share similar features [90]. Training CNNs to learn the similarities achieves the goal of learning contextual semantics. In addition, in events such as ball games, the features of frames representing a batting action should also be smooth, i.e. temporal continuous [200]. Second, frames representing similar motions such as cycling should share similar visual features [201]. More generally, similar objects should share similar motions, which can be learned by CNNs [202]. For instance, similar human poses should also share similar motions [203, 204]. Third, frames representing actions should occur in a certain temporal order. This idea has led to the development of CNNs which learn whether a sequence of frames is in the correct order or not [205, 206].

Imaging data with multiple modalities can be easily used for self-supervised learning. The cross-supervision mentioned above is an obvious strategy to use for multi-modal imaging data. For instance, cameras at different angles offer different views. A siamese CNN could be trained to predict camera poses [88]. More generally, images with the same ego-motion are likely to share similar features which can be learned by CNNs [207]. For videos with audio, it is reasonable to assume similar events share similar audio sound [208]. Exceptionally, in news broadcast videos, similar lip poses represent similar readings [209].

In medical imaging, patients often have follow-up scans. Recognizing scans of the same patient is a

good method of self-supervised learning. Jamaludin et al. [210] proposed a siamese CNN to recognize patients' MR scans and predict the level of vertebral bodies. A large number of scans was collected to train the CNN to recognize MR scans. Therefore, a small number of annotated scans is required for disease prediction. The above approach is one of the first works on self-supervised learning in medical imaging.

Our work also relates to the work of [85], which proposed to combine multiple self-supervised learning tasks to improve the feature learning. In this work, patch relative position prediction [4], colourization [86], exemplar learning [199], and motion segmentation [89] were unified into one architecture. A novel input harmonization method was proposed to enable end-to-end training. Features learned in the individual tasks were then fused with an L1 penalty loss so that their combination could be sparse. The results showed that multi-task self-supervised learning improves subsequent tasks more than single-task self-supervised learning. The disadvantage of multi-task self-supervised learning is the training requires significant computational resources, i.e. 64 GPUs for approximately 16.8K GPU hours.

6.3 Self-supervision Based on Context Restoration

We propose a novel strategy for self-supervised learning which we term *context restoration*. We first introduce this concept before we provide further details of the training process.

6.3.1 Context Restoration

There are two steps in self-supervised learning based on context restoration: generating paired input/output images for training and learning a mapping between them. Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ consisting of N images with no annotations, a new dataset

$$\tilde{\mathcal{X}} = f(\mathcal{X}) \tag{6.1}$$

is generated. Here $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_N\}$. $f(\cdot)$ is a function corrupting the context of original images. Subsequently, a CNN is learned to approximate the function $g(\cdot)$ which is designed to model the mapping $\tilde{\mathbf{x}}_i \mapsto \mathbf{x}_i$, i.e.

$$\mathbf{x}_i = g(\tilde{\mathbf{x}}_i) = f^{-1}(\tilde{\mathbf{x}}_i), \tag{6.2}$$

where $i = 1, 2, \ldots, N$.

Given an image $\mathbf{x}_i$, we randomly select two isolated small patches in $\mathbf{x}_i$ and swap their context. Repeating this process for T times results in $\tilde{\mathbf{x}}_i$. Figure 6.2 demonstrates this process on exemplar images and Algorithm 1 summarises the process in detail. Subsequently, $g(\cdot)$ aims to restore the context using CNN model by learning to approximate $f^{-1}(\cdot)$. This is illustrated in Figure 6.3.

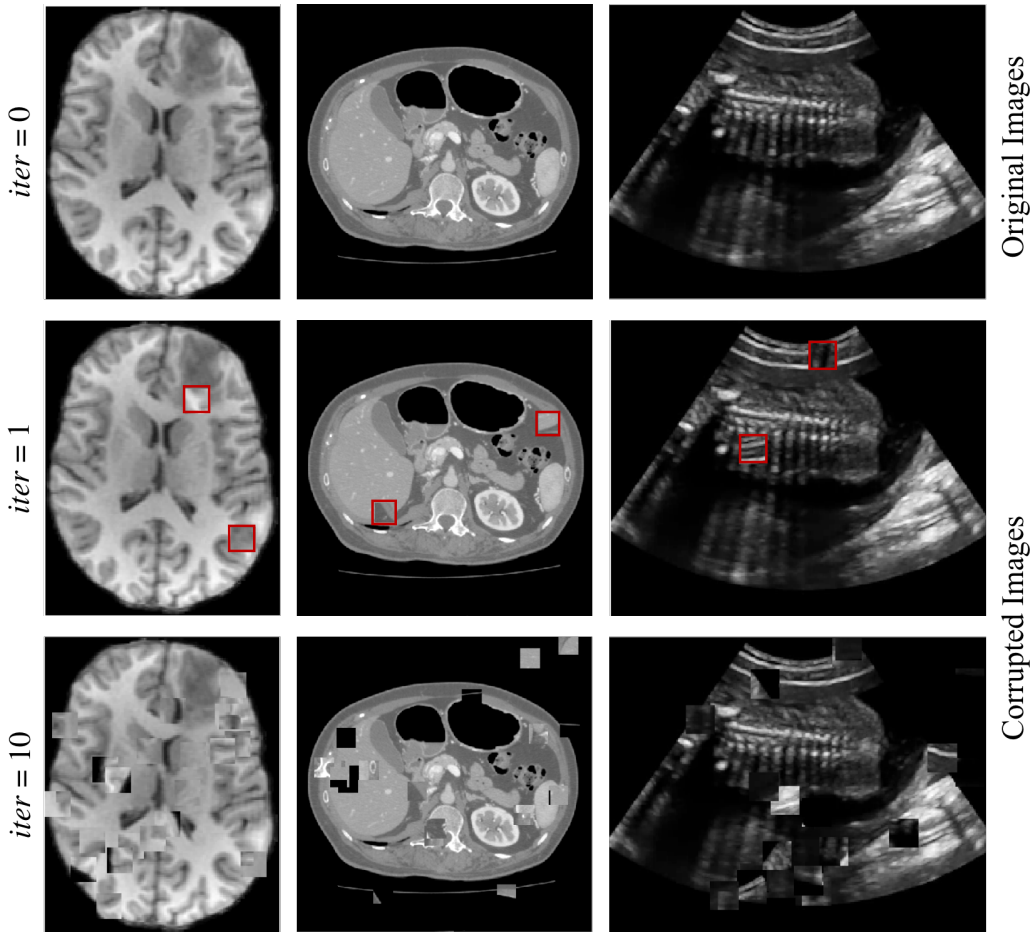


Figure 6.2: Generating training images for self-supervised context disordering: Brain T1 MR image, abdominal CT image, and 2D fetal ultrasound image, respectively. In figures in the second column, red boxes highlight the swapped patches after the first iteration.

Inspired by existing self-supervised learning strategies, a good self-supervised learning strategy should exhibit three key features: 1) features learned in the self-supervised training stage should be meaning-

Algorithm 1: Image context disordering

Input: original image $\mathbf{x_i}$
Output: image with disordered context $\tilde{\mathbf{x}}_\mathbf{i}$
for $iter = 1, 2, \ldots, T$ **do**
  randomly select a patch $p_1 \in \mathbf{x_i}$
  randomly select a patch $p_2 \in \mathbf{x_i}$
  $p_1 \cap p_2 = \emptyset$
  swap p_1 and p_2

ful; 2) self-supervised pretraining is useful for different types of subsequent tasks; and 3) the implementation should be simple. Our proposed context restoration method features all these advantages. For many common problems in medical imaging such as classification, localization, and segmentation, learning image context is key. Therefore, learning the context of images in the self-supervised pretraining stage benefits the subsequent tasks. Restoring the image context can learn image context. Specifically, given the corrupted image $\tilde{\mathbf{x}}_i$, the $g(\cdot)$ function learns to restore it by solving two subtasks: 1) recognising which parts of the image contain corrupted context; 2) reconstructing the correct image context in these areas. Second, the proposed context restoration pretraining is applicable for different types of subsequent tasks by adjusting CNN architecture according to that of subsequent task. Finally, the implementation of the context restoration task is simple and straightforward.

6.3.2 Network Architectures

We model the proposed self-supervised learning strategy – context restoration – using CNNs. The CNNs can be implemented using various different architectures. Most of these networks are image-to-image networks consisting of two parts: an analysis part and a reconstruction part. Figure 6.3 shows an overview of the general architecture of feasible CNNs. The analysis part encodes input disordered images into feature maps and the reconstruction part uses these feature maps to produce output images in correct context.

Analysis Part: The analysis part consists of stacks of convolutional units and downsampling units, extracting feature maps from the input images. The convolutional units can be single convolution layers, residual convolution layers [10], inception layers [211], densely connected convolution layers [11] and so on. The downsampling units could be single pooling layers or inception pooling layers

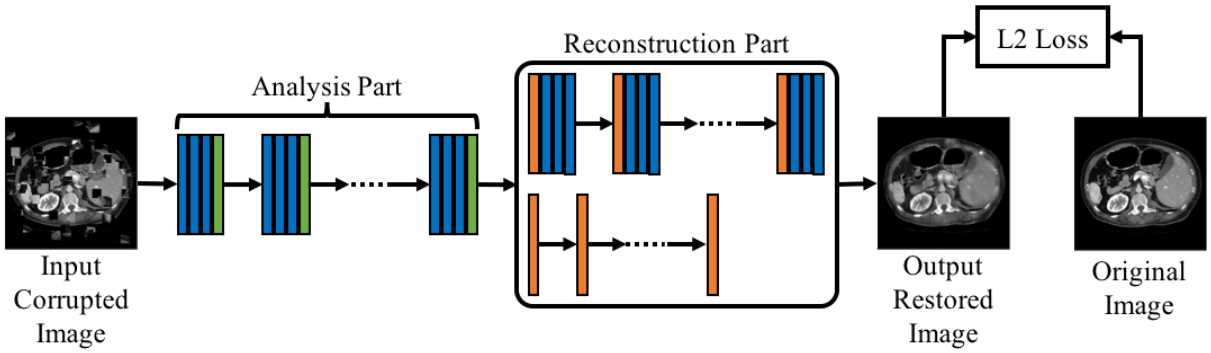


Figure 6.3: General CNN architecture for the context restoration self-supervised learning. In the figure, the blue, green, and orange strides represent convolutional units, downsampling units, and upsampling units, respectively. In the reconstruction part, CNN structures could vary depending on subsequent task type. For subsequent classification tasks, the simple structures such as a few deconvolution layers (2nd row) are preferred. For subsequent segmentation tasks, the complex structures (1st row) consistent with the segmentation CNNs are preferred.

[15, 211] and so on. The CNN weights learned in this part are then used to initialise the subsequent tasks.

Reconstruction Part: The reconstruction part consists of stacks of convolutional layers and upsampling layers, producing output images in which the context information has been restored. The upsampling layers can be deconvolution layers or other upsampling layers. Again, the CNN architectures used here are flexible. In subsequent classification tasks, the CNN weights learned in this part are not used. As suggested by [85], simple CNN layers with a few deconvolution layers are sufficient (see Figure 6.3). In this condition, the analysis part makes most contributions to the context restoration. Therefore, the feature maps resulting from the analysis part are more useful. In subsequent segmentation tasks, the CNN weights learned in this part are then used. In this situation, the CNN architectures of the self-supervised learning and the subsequent main task learning can be consistent. As a result, almost all the weights of the subsequent segmentation CNN can be initialised using those learned in the self-supervised learning. This results in better segmentation results.

Loss Function: We propose to use the L2 loss for training the CNNs for the task of context restoration. As suggested by [5], the L2 loss is sufficient for feature learning although the outputs from context restoration outputs may be blurry.

Implementation: In this work, the CNNs for context restoration employ single convolution layers as the convolutional units. In the analysis part, the architecture is similar to that of the VGG-Net [64],

where there is a pooling layer following a few convolution layers. In the reconstruction part, if the subsequent task is a classification task, then there are only a few deconvolution layers; if the subsequent task is segmentation, then the reconstruction part is in symmetry with the analysis part with concatenation connections, which is similar to a U-Net architecture [6]. The loss function of CNNs in the subsequent tasks is the cross-entropy function.

All the CNNs use the Adam method [133] for optimizing the loss function. We use $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$. The learning rates varies for the different problems. Batch normalization [132] is utilized in all CNNs. Random weights are used for initialization and sampled from a truncated normal distribution with standard deviation of 0.01. The kernel size of the convolution and deconvolution layers is 3×3. The stride size of the convolution layers is 1 and that of the deconvolution layers is 2.

The CNNs implemented in this chapter use the Tensorflow[1] platform. Our experiments are performed on a desktop PC with an Core i7-3770 processor and 32GB RAM and with an NVIDIA TITAN XP GPU processor.

6.4 Experiments and Results

To evaluate the proposed self-supervision approach we have conducted four sets of experiments: First, we show the proposed self-supervision using context restoration task can be performed by CNNs on three different datasets, including brain MR images, abdominal CT images, and fetal US images. In addition, we use the pretrained CNNs for subsequent tasks such as classification, localization, and segmentation, respectively. For each of these problems, a different dataset is used. More importantly, we compare different self-supervised learning strategies, namely, training an auto-encoder [3], self-supervision using patch relative position prediction [4], self-supervision using local context prediction [5], and the proposed context restoration. For each dataset, the self-supervised learning is based on the whole training set. The subsequent tasks are based on the whole, half, and quarter of the training set, respectively.

[1] https://www.tensorflow.org/

6.4.1 Context Restoration Results

We evaluate the CNNs employed for context restoration on three different datasets, including brain MR images, abdominal CT images, and fetal US images. Figure 6.4 shows examples of the three datasets. In all cases, the image context restoration achieve qualitatively good results. A shortcoming is that the L2 loss results in image blur.

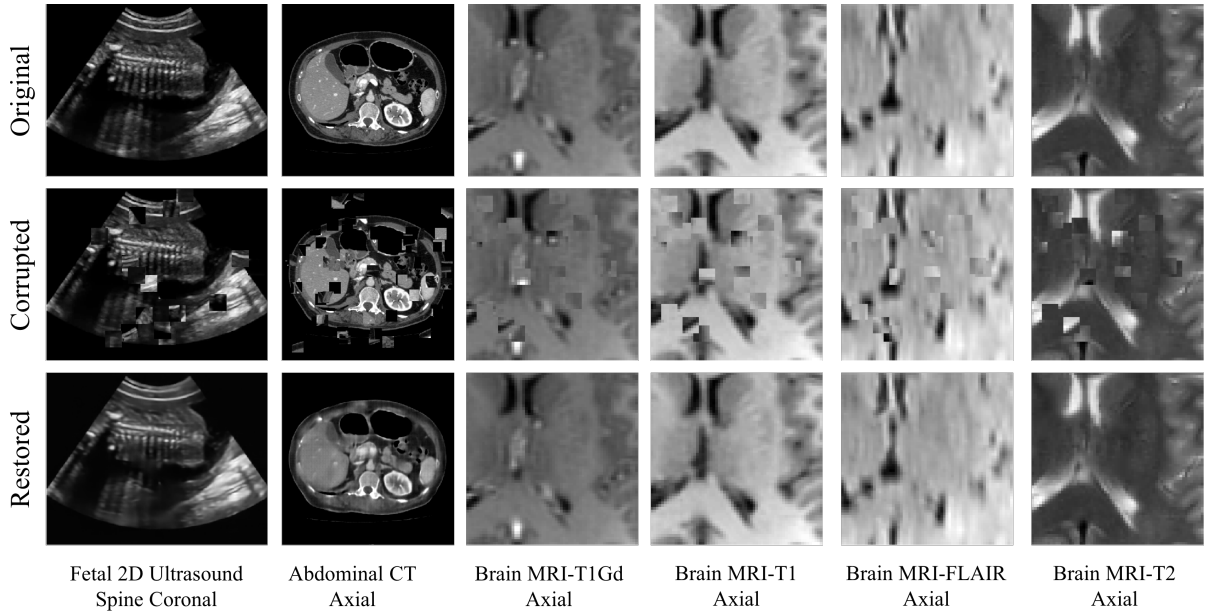


Figure 6.4: Self-srpervision using context restoration: For brain MR images, our training is on 2D image patch level. Therefore, the context restoration is also based on patches.

6.4.2 Fetal Standard Scan Plane Classification

Overview: 2D US imaging is the most widely used medical imaging modality to assess the health of the fetus. In the UK, the fetal abnormality screening programme (FASP) handbook [212] defines guidelines for selecting a number of standard scan planes, which are used to make biometric measurements and possible abnormalities. However, US images often have low quality because of noise, artefacts, shadows, etc. Therefore, interpreting fetal US images is challenging. Baumgartner et al. proposed a novel CNN-based approach (known as the SonoNet) to detect and localise the defined 13 different standard scan planes in real-time from US images [213].

Dataset: We use the same dataset as used in [213]. Our dataset consists of 2694 2D ultrasound

examinations of fetuses with gestational ages between 18 and 22 weeks. More details about the image acquisition protocol can be found in [213]. Figure 6.5 shows examples of each class of scan planes.

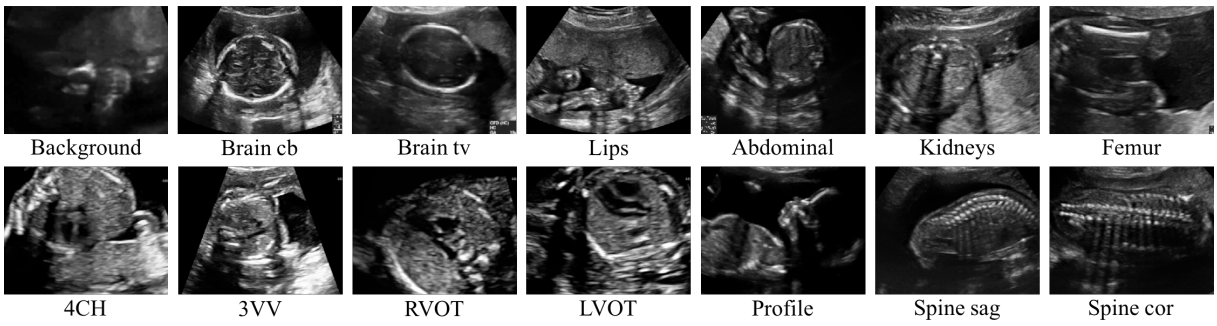


Figure 6.5: Examples of standard scan planes and background views of 2D fetal ultrasound images. The standard scan planes consist of brain view at the level of the cerebellum (Brain cb), brain view at posterior horn of the ventricle (Brain tv), coronal view of the lips and nose (Lips), standard abdominal view at stomach level (Abdominal), axial kidneys view (Kidneys), standard femur view (Femur), sagittal spine view (Spine sag), coronal spine view (Spine cor), four chamber view (4CH), three vessel view (3VV), right ventricular outflow tract (RVOT), left ventricular outflow tract (LVOT), and median facial profile (Profile).

Implementation: The CNN for this classification problem is the SonoNet-64 which achieved the best performance in [213]. In terms of the training strategy, we use a fixed learning rate of 0.01. In the original training, each batch consists of 2 images from each of the standard scan plane categories and 26 images from background images. As a multi-class classification problem, the numbers of instances across classes are imbalanced. In our implementation, we sample equal number of frames for each class, including the background.

Evaluation: As in [213], we evaluate the performance of CNNs in this classification task using the precision, recall, and the F1-score.

Results: Table 6.3 displays the results of performance of the CNNs under different configurations. Balancing the numbers of instances in each class significantly improves the performance in all three metrics.

In training in random initialisation situations, it is not surprising that less training data leads to worse results. When the SonoNet is trained on half of the training data, the precision and recall both decrease, which lead to the decrease of the F1-score. Interestingly, when the SonoNet is trained on

Table 6.3: The classification of standard scan planes of fetal 2D ultrasound images. The entries in bold highlight the best comparable results.

Training data%	Initialisation	Precision (%)	Recall (%)	F1-score (%)
100% [213]	Random	80.60	86.00	82.80
100%, Ours	Random	**89.39**	**89.66**	**89.42**
50%	Random	84.69	84.94	84.64
	Auto-encoder [3]	84.63	86.09	84.50
	Relative positions [4]	85.15	86.79	84.74
	Context prediction [5]	84.43	85.27	84.43
	Context restoration	**85.52**	**87.56**	**85.94**
25%	Random	57.23	78.99	62.85
	Auto-encoder [3]	55.54	82.87	62.32
	Relative positions [4]	61.01	83.09	66.38
	Context prediction [5]	57.73	81.58	63.10
	Context restoration	**65.69**	**85.25**	**69.93**

quarter of the training data, the precision decreases significantly while there is only slight decrease in terms of the recall. This suggests a large number of false positives (FPs) occur.

With the help of self-supervised pretraining, the performance of CNNs when using small training sets can be improved. Specifically, when learning on half of training images, the F1-scores keep stable in most cases except where the SonoNet is pretrained based on context restoration. In this scenario, the baseline (i.e. random initialisation) is not far away from the ceiling (i.e. SonoNet on the whole training set). Therefore, it is difficult to obtain improvements. The SonoNet pretrained using context restoration can only offers marginal improvement. When learning using only a quarter of training images, the SonoNet with feature initialisation from the auto-encoder pretraining still cannot improve the baseline; while SonoNets using other pretraining strategies perform better than the baseline. Our context restoration pretraining improves the SonoNet performance the most. This suggests that context restoration pretraining is more useful for image classification in this case.

6.4.3 Abdominal Multi-organ Localization

Overview: In many medical image analysis problems, localization anatomical structures is a prerequisite. For instance, in the liver segmentation challenge [214] hosted in MICCAI 2007, the provided

CT images were cropped such that the livers were roughly localized. This excludes irrelevant organs and tissue and benefits the segmentation. However, manual cropping requires expert knowledge and costly. de Vos et al. [215] proposed a novel approach which can localize anatomical structures in 3D medical images. This approach defines the localization as discovering bounding boxes in 3D images so that regions within these bounding boxes contain target anatomical structures (see Figure 6.6). Following this idea, we localise multiple abominal organs in CT images. The organs of interest are pancreas, kidneys, liver, and spleen.

Dataset: A dataset of 3D abdominal CT image from 150 subjects is employed. The patient demographics and image acquisition details can be found in [153]. We normalize the volume intensities in zero mean and unit deviation before analysis. The whole dataset is randomly divided into two equal halves. The first half is used for training and validation and the other half is used for testing. Images in this dataset were annotated at voxel level. We derive the reference bounding boxes and slice labels (organ presence) using these annotations.

Implementation: The CNN for multi-organ localization task is similar to the SonoNet [213]. It has one more stack of convolution and pooling layers than the SonoNet since the input images are 512×512 which is approximately twice larger than the processed 2D ultrasound frames in each side. The CNN for localization is also equipped with a global mean pooling layer. The output of this CNN is a prediction vector with K elements indicating the probabilities of presence of the K organs. The learning rate in this task is fixed as 0.001.

Evaluation: We follow [215] that distances (in mm) from the reference bounding boxes to the predicted bounding boxes are used to evaluate the localization performance. Specifically, we compute the distances of the centroids and walls between bounding boxes.

Results: Table 6.4 displays localization performance of the CNN in different training strategies. Initialising by pretrained features, particularly those from context restoration tasks, improves the CNN performance.

Performance is compared among CNNs using different pretraining strategies. Training on incomplete training set using random initialization is used as baseline in each comparison group. Within each

Table 6.4: The performance of the CNN solving the multi-organ localization problem in different training settings. The entries in bold highlight the best comparable results. The RD, AE, RP, CP, CR are short for random, auto-encoder [3], relative positions [4], context prediction [5], and our proposed context restoration. The numbers displayed are the mean ± std distances in mm.

Train data %	Init.	Left Kidney		Right Kidney	
		Centroid	Wall	Centroid	Wall
100%	RD	6.45 ± 8.47	3.68 ± 21.41	5.71 ± 10.17	2.79 ± 23.65
50%	RD	17.49 ± 49.67	9.36 ± 75.00	10.40 ± 30.37	5.89 ± 48.28
	AE	12.79 ± 38.67	6.84 ± 56.97	20.44 ± 41.48	11.52 ± 67.01
	RP	12.11 ± 39.01	6.75 ± 61.67	10.61 ± 30.41	5.77 ± 48.64
	CP	11.95 ± 38.97	6.82 ± 61.23	8.30 ± 11.92	4.47 ± 27.83
	CR	$\mathbf{5.99 \pm 9.83}$	$\mathbf{3.16 \pm 22.66}$	$\mathbf{5.83 \pm 10.10}$	$\mathbf{2.90 \pm 22.04}$
25%	RD	28.23 ± 71.95	15.87 ± 107.18	12.71 ± 30.39	6.77 ± 49.26
	AE	25.90 ± 65.64	14.40 ± 98.28	36.28 ± 73.65	19.55 ± 111.46
	RP	27.65 ± 75.31	15.41 ± 111.82	$\mathbf{8.34 \pm 11.22}$	$\mathbf{3.97 \pm 23.26}$
	CP	21.86 ± 60.28	13.03 ± 90.92	15.58 ± 35.3	8.42 ± 57.53
	CR	$\mathbf{7.63 \pm 9.02}$	$\mathbf{3.94 \pm 22.78}$	17.51 ± 52.67	9.8 ± 78.57

Train data %	Init.	Pancreas		Liver		Spleen	
		Centroid	Wall	Centroid	Wall	Centroid	Wall
100%	RD	13.39 ± 9.73	8.98 ± 23.27	7.50 ± 5.22	4.35 ± 14.07	6.63 ± 9.68	4.10 ± 23.02
50%	RD	16.45 ± 9.00	10.74 ± 26.77	12.79 ± 8.19	6.89 ± 22.6	13.24 ± 36.97	8.54 ± 56.87
	AE	15.59 ± 8.51	10.35 ± 24.35	14.07 ± 8.66	7.41 ± 24.39	12.36 ± 11.31	8.54 ± 31.16
	RP	15.54 ± 7.98	11.13 ± 23.50	10.12 ± 8.85	6.18 ± 22.31	7.64 ± 10.16	4.77 ± 24.41
	CP	14.76 ± 8.78	$\mathbf{10.07 \pm 26.26}$	9.91 ± 6.78	5.03 ± 15.39	7.79 ± 11.41	4.82 ± 25.98
	CR	$\mathbf{14.76 \pm 8.10}$	10.14 ± 24.86	$\mathbf{8.91 \pm 6.20}$	$\mathbf{4.67 \pm 16.83}$	$\mathbf{7.07 \pm 9.54}$	$\mathbf{4.05 \pm 22.17}$
25%	RD	22.09 ± 11.72	17.14 ± 39.23	12.02 ± 6.46	7.14 ± 20.27	24.86 ± 36.64	15.30 ± 61.38
	AE	17.67 ± 8.40	12.24 ± 25.54	16.79 ± 9.47	9.56 ± 28.30	22.65 ± 47.91	13.95 ± 73.05
	RP	17.84 ± 8.94	$\mathbf{11.74 \pm 25.06}$	15.59 ± 9.79	9.25 ± 29.74	14.51 ± 38.89	9.95 ± 62.12
	CP	21.81 ± 11.44	18.59 ± 41.57	11.40 ± 8.69	$\mathbf{6.18 \pm 22.50}$	10.34 ± 9.92	7.56 ± 27.58
	CR	$\mathbf{16.01 \pm 8.46}$	11.78 ± 28.79	$\mathbf{11.17 \pm 9.03}$	7.52 ± 25.68	$\mathbf{8.39 \pm 6.28}$	$\mathbf{5.82 \pm 19.50}$

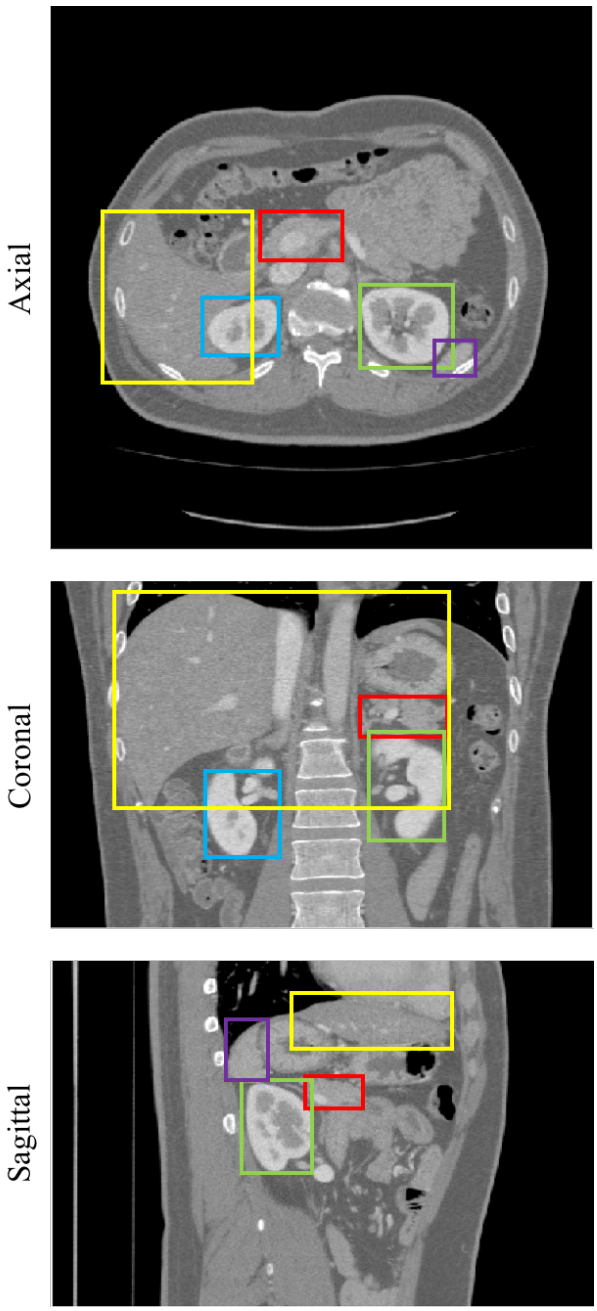


Figure 6.6: An example of abdominal CT image in axial, coronal, and sagittal views. The pancreas, left kidney, right kidney, liver, and spleen are colours in red, green, blue, yellow, and purple, respectively.

group, the CNN pretrained using the auto-encoder sometimes improves the performance upon the baseline. For instance, on half training data, it improves the centroid prediction of pancreas. However, it is worse than the baseline in terms of liver. In total, the results cannot verify auto-encoding pretraining improves the CNN performance. In contrast, pretraining based on relative position prediction and context prediction improves the CNN performance. Specifically, in most cases, pretraining of these two tasks decreases the errors on baselines in terms of both centroid and walls. Importantly,

pretraining based on context restoration results in more localization improvements. In some cases, the CNN using context restoration pretraining is comparable to or even better than none pretraining on more annotated training data. For instance, in terms of left kidney, the CNN on half training data slightly outperforms that on all the training data; in terms of spleen, the CNN on a quarter training data performs better than the one on half training data. These improvements cannot be achieved by CNNs using other pretraining strategies.

In terms of different organs, the distance variance of centroid and walls in kidneys is significantly larger than that of other organs. This is because not all patients have two kidneys. It is challenging for CNNs to distinguish two kidneys individually because of inter-subject variance. CNNs are more likely to make mistakes based on less training data. Although the CNN pretrained using the RP method on quarter training data outperforms that using context restoration pretraining, it performs much worse in left kidney. Regarding the pancreas, the performance of CNNs without pretraining decreases slightly when the training data halves. However, it decreases significantly when there is only quarter training data. In the opposite, in terms of the liver, the CNN performance decreases sharply with half training data; while it remains stable with quarter training data. On the spleen, the situation is different. The CNN performance keeps decreasing rapidly with less and less training data. It is noteworthy that if less training data leads to significant decrease of results, self-supervised learning is likely to improve the results significantly.

6.4.4 Brain Tumour Segmentation

Overview: Gliomas are the major brain tumours occurring in adults. They are routinely assessed using MR imaging [27]. Accurate segmentation of gliomas on MR image is a key step for quantification. Our segmentation task is based on the Brain tumour segmentation (BraTS) chanllenge [162]. The task is to segment the necrotic and non-enhancing tissues, the peritumoral edema, and gadolinium enhancing tissues of tumour [161] on multi-modal MR images. Figure 6.7 shows such an example.

Dataset: We use the dataset of the BraTS 2017 challenge which consists of 285 subjects. Each subject has MR images in multiple modalities, namely, native T1 (T1), post-contrast T1-weighted

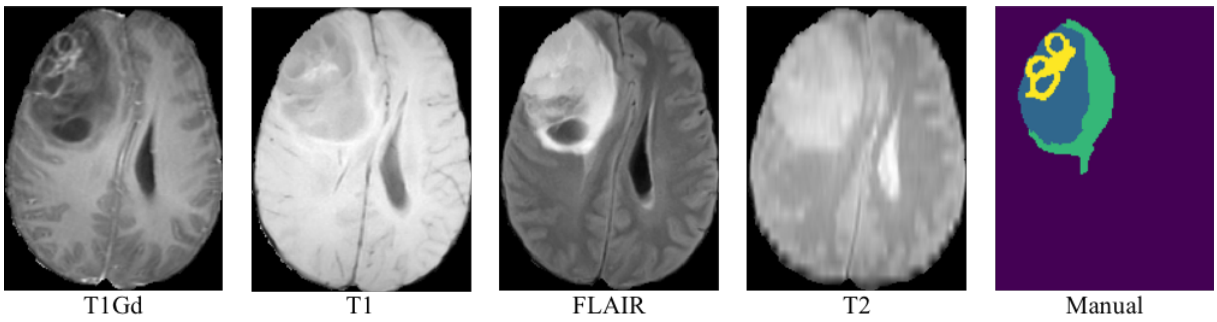


Figure 6.7: An example of MR image in multiple modalities with gliomas and the tumour structure annotations. In the manual annotation image, the background, edema, non-enhancing tumours, and enhancing tumours are coloured in purple, green, blue, and yellow, respectively.

(T1-Gd), T2-weighted (T2), T2 fluid attenuated inversion recovery (FLAIR). These images were preprocessed that images in different modalities are co-registered into the same anatomical template; skulls are removed; and voxels are resampled into the isotropic resolution ($1mm^3$) [162]. Intensities are normalized to zero mean and unit variance. We use 142 out of the 285 images for training and validation and remaining 143 ones for testing.

Implementation: For the tumour segmentation in this work, we use a 2D patch-based CNN approach as suggested in [24,185] in medical image segmentation since medical images usually have large sizes while lesions of interest are small. Figure 6.4 shows an example of such patches. The patch size used is 64×64. The CNN used in this experiment is a 2D U-Net [6]. The learning rate is fixed as 0.001. We follow the post-processing strategy proposed in [179]: a 3D dense conditional random fields (CRFs) [216] is used to refine the output of whole tumour structures; isolated voxel clusters of whole tumours less than 1000 voxel size are then removed based on the connected component analysis; the predicted voxels of tumour cores outside the regions of whole tumours are removed.

Evaluation: The evaluation is not based on three tumour classes individually. It is based on the following three classes: the whole tumour region which include all tumour structures, the tumour core region which include tumour structures except edema, and the enhancing tumour core region. We use the same evaluation metrics in the BraTS 2017 challenge: Dice score, sensitivity, specificity, and Hausdorff distance. Particularly, we use a robust version of the Hausdorff distance (Hausdorff95), which measures the 95% quantile, instead of the maximum distance between two surfaces.

Results: Table 6.5 shows the results on the BraTS problem. The general experiment settings are sim-

ilar to the previous experiments. According to the results, U-Nets [6] initialised by context restoration pretraining achieve the best performance in total.

In terms of different pretraining strategies, the auto-encoding pretraining does not improve CNN performance, which has been verified in previous experiments. This is also similar to the previous experiments that pretraining based on relative positions and context prediction tasks improves the segmentations but they are not as good as the pretraining based on the context restoration task. Again, self-supervision based on context restoration offers best pretraining startegy for the segmentation task.

The decrease in U-Net performance is not significant every time when the size of the training data halves. Therefore, the differences in performance among different self-supervision strategies are not significant. The performance using self-supervision based on context restoration approaches that of random initialisation on a larger dataset. For instance, using 50% of the training set, the proposed self-supervision strategy offers similar performance to using the whole training set. The Dice score in enhanced tumour core, the sensitivity in non-enhanced and enhanced tumour cores, and the Hausdorff distances in all aspects are even slightly better.

6.5 Discussion and Conclusion

In this work, we proposed a novel self-supervised learning strategy based on context restoration. This enables CNNs to learn useful image semantics without any labels. The subsequent task-specific CNNs benefit from this pretraining. We conclude from the existing self-supervised feature learning literature that the ideal pretraining task should have similar goal to the subsequent task. Particularly, in medical image analysis, the image context is the common feature for classification, localization/detection, and segmentation tasks. Therefore, the context restoration learning contribute to learning features for these goals.

In addition, the CNNs for context restoration can be structured in flexible architectures depending on subsequent tasks. The idea is to ensure subsequent tasks can make full advantages of the weights from pretrained CNNs. Furthermore, the implementation of the context restoration task is simple and

Table 6.5: The segmentation results of the customised U-Nets [6] in different training settings. The entries in bold highlight the best comparable results. The RD, AE, RP, CP, CR are short for random, auto-encoder [3], relative positions [4], context prediction [5], and our proposed context restoration.

Train data %	Init.	Dice %			Sensitivity %			Specificity %			Hausdorff95		
		Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.
100%	RD	86.56	77.04	66.31	87.05	77.28	77.62	99.88	99.94	99.95	30.78	25.03	25.74
50%	RD	84.41	75.55	65.11	**84.75**	77.76	80.2	99.86	99.91	99.94	31.29	25.26	26.81
	AE	84.33	71.85	65.07	84.71	74.19	77.38	99.87	99.91	**99.95**	33.36	25.24	24.56
	RP	84.38	75.65	66.73	84.65	77.02	79.48	99.87	**99.92**	**99.95**	36.43	23.15	**20.69**
	CP	84.54	73.86	66.01	84.59	75.28	79.46	99.86	**99.92**	99.94	33.59	28.59	26.90
	CR	**85.57**	**76.2**	**68.24**	83.83	**78.17**	**80.53**	**99.89**	**99.92**	**99.95**	**26.41**	**20.34**	24.38
25%	RD	81.91	71.22	62.57	84.08	75.68	75.98	99.82	99.89	**99.94**	36.34	37.21	31.57
	AE	83.05	68.92	61.28	83.90	76.52	76.75	**99.85**	99.86	99.93	33.21	34.9	31.95
	RP	82.38	71.33	61.86	84.23	72.53	75.38	99.83	**99.92**	**99.94**	37.83	**31.81**	31.04
	CP	83.19	71.55	62.77	**85.75**	73.68	76.88	99.83	99.91	**99.94**	36.21	36.45	31.90
	CR	**84.27**	**73.43**	**64.12**	85.57	**78.79**	**79.14**	**99.85**	99.89	**99.94**	**33.15**	32.18	**30.61**

straightforward, meaning that it can be widely used. Compared with the existing strategies such as relative positions and context prediction, solving the context restoration task requires pattern recognition and prediction, which ensures the context restoration task offers more efficient image semantics.

We have validated the proposed context restoration pretraining on three types of representative tasks in medical image analysis, which are classification, localization, and segmentation. Each of these tasks are based on a different type of medical images. The classification task is based on fetal 2D ultrasound images; the localization task is based on abdominal CT image; and the segmentation task is based on multi-modal brain MR images. In all three tasks, context restoration pretraining outperforms other pretraining methods. These results underlines the advantages of our context restoration strategy. In our experiments, we found that if the reduction of training data causes significant performance decrease, the context restoration pretraining can offer significant performance improvement over the baselines.

In computer vision, many CNNs are pretrained before the main task. For instance, the Faster R-CNN [195] is based on the pretraining of the VGG-Net [64]. This type of pretraining leads to good detection results in the Faster R-CNN. However, it was reported that the self-supervised pretraining is not as good as the supervised pretraining [217]. This is not verified in this work since in medical image analysis, it is difficult to conduct supervised pretraining, which requires a large number of annotations. However, it is noteworthy to exploring more powerful self-supervised learning method so that the self-supervised pretraining can be as good as supervised pretraining in the future. Furthermore, comprehensive image augmentation contributes to the model performance based on small training datasets. We plan to study the effect of data augmentation when using self-supervised learning.

Chapter 7

Small Vessel Disease Identification on CT Images

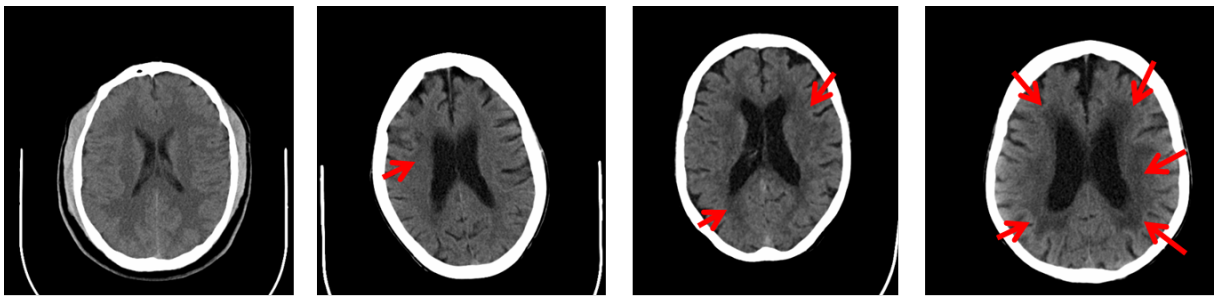
The work in this chapter is based on:

- L. Chen, T. Tong, C.P. Ho, R. Patel, D. Cohen, A.C. Dawson, O. Halse, O. Geraghty, P.E. Rinne, C.J. White, T. Nakornchai, P. Bentley, and D. Rueckert, "Identification of cerebral small vessel disease using multiple instance learning," in *Proceedings of the International Conference of Medical Image Comupting and Computer-Assisted Intervention*, pp. 523–530, 2015.

7.1 Introduction

Fazekas et al. [218] proposed a standard approach for SVD grading. In this approach, SVD is divided into four categories according to the degree of the lesion severities: absent, mild, moderate, and severe. Generally, mild SVD is associated with normal brain ageing while moderate or severe SVD suggests potential risks for diseases such as stroke. It can be seen in Figure 7.1 that the lesion severity relates to lesion volumes and if it extends to gray matter (GM). As such, if lesion volumes and positions can be quantified, then Fazekas grading can be addressed using classic machine learning algorithms, including logistic regression, SVM, and RF. However, the SVD lesion quantification is

challenging so Fazekas scores are qualitatively graded by experts. Lesion quantification requires fine annotations at voxel level, which is expensive in terms of time and human resource. In addition, it is difficult to distinguish SVD lesions and normal tissue since the lesions look blurry, particularly at boundaries.



(a) Normal (b) Mild (c) Moderate (d) Severe

Figure 7.1: Examples of CT images of the brain: (a) normal brain appearance, (b) brain with mild cerebral SVD, (c) brain with moderate SVD, and (d) cerebrum with severe SVD. The red arrows point out where the lesions are.

There have been a large number of studies focusing on the automatic analysis of brain MR images. For instance, in terms of Alzheimer's Disease (AD), machine learning techniques have been extensively used to classify controls and patients [219–221]. However, there are very few works that focus on the classification of subjects suffering from stroke and even fewer which use CT images [222]. The cutting-edge studies on CT images [223–225], are typically based on statistical values and threshold. These methods are fairly simple and cannot perform well on large datasets. To the best of our knowledge no machine learning approach has been proposed for the identification of SVD in a large dataset of CT images. This is because CT images are usually not annotated at voxel level so the key features such as lesion volumes cannot be achieved. Classic machine learning algorithms are not applicable.

In the context of similar challenging classification problems in medical images, weakly supervised machine learning approaches, particularly MIL [226] have been very successful. The reason is that the diagnosis of medical images is usually based on some ROIs instead of the whole image; however annotations on these ROIs are usually unavailable and annotations are only on the image level. Conventional machine learning methods analyzing the whole images are difficult to achieve satisfactory classification results. In contrast, the MIL can solve classification problems where annotations are coarse-grained. In MIL, each image can be recognized as a bag containing a number of instances.

The instances are features extracted from the image. Different images have different number of instances. In binary classification of MIL, a bag is positive if there is one positive instance in it; otherwise the bag is negative. The labels of the instances are unknown but the label of the image is known. For instance, Tong et al. [96] employed MIL to classify subjects in the context of AD. In this case, only subject level classes are available but the class of a subject is strongly associated with the regions around the hippocampus. To this end, the image patches around the hippocampus were extracted and packed into bags. Then the SVM classifier was used on the bag level to achieve desirable classification results.

There are many MIL methods that have been developed and applied, e.g. MIS-Boost [93], MIForest [227], and EM-Diverse Density [228]. The MIS-Boost proposed by Akbas et al. [93] is based on boosting. It outperforms a number of other similar algorithms on several benchmark datasets. This approach aims to learn a specific instance for each weak classifier, which is able to discriminate two categories of bags.

In this chapter, we address the problem of automatic SVD identification on CT images. An MIL-based framework is proposed to classify SVD into normal (absent and mild) and abnormal (moderate and severe) groups on a large dataset of CT images. In our approach, each CT image is regarded as a bag, which contains a number of image patches (i.e. instances). The MIS-Boost algorithm is formulated to apply on the bags for classification. We achieve good results distinguishing SVD lesions with different severity groups. Comparisons among other state-of-the-art algorithms show the advantages of our method.

7.2 Methods

7.2.1 Overview

In our approach, we first build feature bags upon images. Instances in these bags are a number of image patches. Afterwards, a classifier based on boosting is formulated to learn several featured instances, which are used to distinguish bags. As such, patients with absent/mild SVD and moder-

ate/severe SVD can be discriminated without fine-grained annotations.

7.2.2 Patch Extraction

In MIL, each bag contains a number of instances, which are patches in our case. Patches were extracted from original CT images since the slice thickness varies between different scans and resampling them to a constant voxel size will reduce the image quality. The extraction was guided by an atlas, which shows the regions with high probability of SVD lesions. In order to construct such an atlas, we collected 277 MR images with SVD. For all these MR images, clinical experts manually outlined ROIs corresponding to the SVD lesions. They were then registered and normalized onto a standard template so that we are able to obtain the lesion atlas. The template was developed by Rorden et al. [229]. The atlas constructed shows the probability for each voxel in the brain to be part of an SVD lesion. We excluded the regions with very low abnormal probability ($< 4\%$) in the atlas since they are likely to be outliers. Finally, the lesion atlas was mapped back to each individual CT image so that for each CT image a native lesion atlas is available which shows regions with high probability of lesions. Figure 7.2 visualizes this processing pipeline.

Details of mapping the lesion atlas to native spaces are as follows. First, all CT images were resampled to a uniform voxel size. Separate image volumes from the same subjects were joined as single volumes. Subsequently, we corrected the gantry tilt and rigidly co-registered all volumes to the template space. Following this step, a non-rigid registration [116] was performed between individual images and the template. Finally, the lesion atlases in individual native spaces were achieved by inverting the deformations.

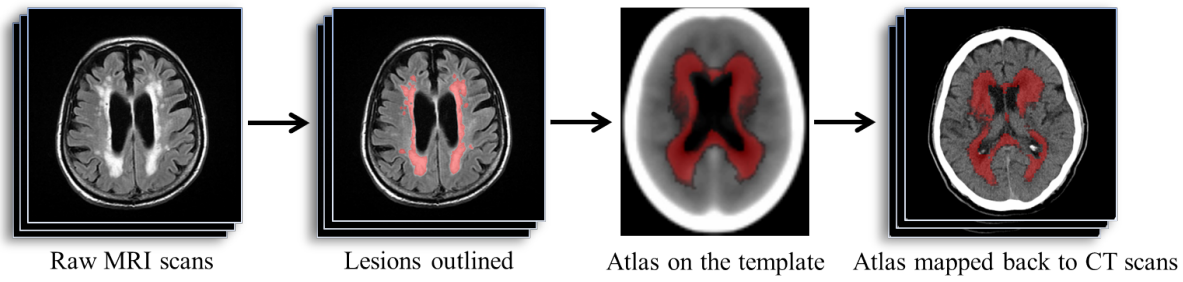


Figure 7.2: The process of atlas construction and mapping back. The red regions are the ROIs for patch extraction.

7.2.3 MIS-Boost

Given bags and their labels, MIL is recognized as a supervised learning method, which learns the mapping $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ is a set of training data and $\mathbf{Y} = \{-1, +1\}$ is the set of corresponding labels. In this case, $\mathbf{X} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_N\}$ and for each bag $\mathbf{B}_i = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_{n_i}\}$, where $\mathbf{I}_k \in \mathbb{R}^r$ is the k-th instance in bag $\mathbf{B}_i$. N is the number of bags. n_i is the number of instances in the i-th bag. r is the size of a patch. The boosting-based MIL proposed in [93] aims to learn a "bag-level" classifier

$$F(\mathbf{B}) = \text{sign}\left(\sum_{m=1}^{M} f_m(\mathbf{B})\right), \tag{7.1}$$

where $f_m(\cdot), m = 1, 2, \ldots, M$, are weak classifiers defined as

$$f_m(\mathbf{B}) = \frac{2}{1 + e^{-(\beta_1 D(\mathbf{p}_m, \mathbf{B}) + \beta_0)}} - 1. \tag{7.2}$$

The task of each weak classifier is to find a patch $\mathbf{p}_m$, which serves as an instance, to discriminate different bags. The distance from an instance to a bag is defined as:

$$D(\mathbf{p}_m, \mathbf{B}) = \sum_{k=1}^{n} \pi_k d(\mathbf{p}_m, \mathbf{I}_k), \tag{7.3}$$

where

$$d(\mathbf{p}_m, \mathbf{I}_k) = \|\mathbf{p}_m - \mathbf{I}_k\|_2, \tag{7.4}$$

and

$$\pi_k = \frac{e^{-\alpha d(\mathbf{p}_m, \mathbf{I}_k)}}{\sum_{l=1}^{n} e^{-\alpha d(\mathbf{p}_m, \mathbf{I}_l)}}. \tag{7.5}$$

$d(\mathbf{p}_m, \mathbf{I}_k)$ is the distance between the specific instance $\mathbf{p}_m$ and the k-th instance $\mathbf{I}_k$ in the bag, which is the standard Euclidean distance. π_k is the associated weight. α is a constant and is set as $1e-4$ in this case. $D(\mathbf{p}_m, \mathbf{B})$ is the weighted average distance of $\mathbf{p}_m$ to each instance in the bag. $f_m(\cdot)$ maps this distance into the range $[-1, 1]$.

In order to learn $\mathbf{p}_m$, an error function is defined based on the Adaboost algorithm [230]. We obtained the parameters β_0, β_1, and $\mathbf{p}_m$ by minimizing the weighted error between the ground truth labels and

the predictions made by weak classifiers.

$$\min_{\mathbf{p}_m,\beta_0,\beta_1} \varepsilon_m = \sum_{i=1}^{N} w_i (y_i - f_m(\mathbf{B}_i))^2 \tag{7.6}$$

In [93] the optimization problem is solved via a coordinate descent algorithm. This uses a line-search method and therefore does not require the calculation of derivatives. However, each iteration is very time-consuming. In this work, we propose to use the region-trust-reflective method [231] to allow a more efficient optimization. Therefore, we formulate optimization of the objective function as a non-linear least square fitting problem.

For initialization, we performed k-means clustering for all instances in all the bags. The k-means algorithm was randomly initialized. The resulting K clustering centres were used as input for the initial $\mathbf{p}_m$ and we selected one leading to the minimum error ε_m among them. The K is set as 3 in our implementation, representing three types of patches: 1) patches with normal tissue, 2) patches with SVD, 3) others. In order to decide on the number of weak classifiers M, we split the training dataset into sub-training and validation sets and pick up the optimal M with minimum validation error. The work-flow of the algorithm is demonstrated in Algorithm 2.

Algorithm 2: Pseudo-code for the MIS-Boost algorithm

Input: training data $\mathbf{X}$ and $\mathbf{Y}$
Initialization
Split $\mathbf{X}$ into $\mathbf{X}_{train}$ and $\mathbf{X}_{valid}$; $\mathbf{Y}$ into $\mathbf{Y}_{train}$ and $\mathbf{Y}_{valid}$
for $m = 1 \dots M$ **do**
  $[\mathbf{p}_m, \beta_{0m}, \beta_{1m}] \leftarrow \arg\min \varepsilon_m$
  $w_i \leftarrow w_i e^{-y_i f_m(\mathbf{X}_{train})}/Z$, Z is the normalization term
  $F \leftarrow F + f_m(\mathbf{X}_{valid})$
  $error_m \leftarrow$ evaluate F on $\mathbf{Y}_{valid}$
$M \leftarrow \arg\min error$
Output: $F(\mathbf{B}) \leftarrow \sum_{m=1}^{M} f_m(\mathbf{B})$

7.3 Experiments and Results

7.3.1 Imaging Data and Pre-processing

In this study, all the data was collected from a local hospital. We collected 627 baseline CT brain images with stroke. For all patients, the imaging was carried out within a short time window after stroke (4.5 hours). The average age of these subjects is 70.75 ± 10.83. There are 326 male and 301 female participants. The labels of these images were assessed by experts according to the Fazekas scoring system [218]. The inter-rater consistency is about 75% among experts.

In order to reduce the radiation burden for patients, in some subjects the brains were scanned in two separate volumes including the cerebrum and the base using different voxel sizes. For the images scanned separately, the voxel sizes of cerebrum and base are approximately $0.45 \times 0.45 \times 7.2$ mm and $0.45 \times 0.45 \times 2.4$ mm, respectively. The voxel size of the whole-brain scans is approximately $0.38 \times 0.38 \times 3$ mm. The template's voxel size is $2 \times 2 \times 2$ mm.

The atlas mapping pipeline failed for 37 CT scans because of poor image quality and/or patient movement. These subjects were excluded and we used the remaining 590 scans in the following experiments, which consists of 350 subjects with absent/mild SVD and 240 subjects with moderate/severe SVD.

7.3.2 Patch-Based Identification of SVD

In order to obtain two SVD groups which are balanced in terms of number of subjects, we randomly sampled 240 subjects from the absent or mild group and performed leave-10%-out cross-validation. The random sampling was repeated for $T = 10$ times and the final results are average values of the T repeats. In this paper, abnormal bags and instances are regarded as positive.

In MIS-Boost, each subject is modelled as a bag, which can contain a number of patches as the instances. The patches were extracted from the region of interest (ROI) according to the atlas. The ROI is defined by those voxels in which the prior probability for lesions is high. As different original

CT scans have different numbers of slices and the size of the brain varies, different bags contain different numbers of instances. We obtain in average 2313 patches (std: 762) in a bag. Given the different slice thickness of the different scans, 2D patches were extracted with a patch size of 15×15. The results are shown in Table 7.1.

Table 7.1: Classification performance of different classifiers and features. Results of MIS-Boost and RF are based on T times cross-validation.

Classifier	Feature	Accuracy(%)	Sensitivity(%)	Specificity(%)
MIS-Boost	Patch in ROI	**75.04**$\pm$1.37	**80.17**$\pm$1.65	69.92 ± 1.37
RF	Voxel in ROI	70.65 ± 0.03	69.63 ± 0.04	**71.67**$\pm$0.04
	Voxel in whole brain	65.25 ± 0.02	65.64 ± 0.04	64.96 ± 0.04
Threshold	t-Score	54.07	5.42	48.64

In order to demonstrate the advantages of our model, we compared the results to those obtained using alternative approaches. We compared our approach to RF [44]. It is one of the most popular standard machine learning methods and has achieved a notable success in classification of AD patients and controls using imaging data [232]. As the CT images have been registered and normalized to the template, the voxels of processed images were selected as features for the RF. Voxels were extracted from the whole brain and the ROIs, respectively.

We also compared the approach by [225] which has shown the ability for automated stroke lesion delineation on brain CT images. In this approach a t-score map is calculated, which can be used to delineate stroke lesions when combined with a carefully selected threshold. Since acute stroke lesions are similar to SVD in terms of intensity and texture, this approach can be tested in terms of its performance for the evaluation of SVD. We collected $N_c = 307$ CT images without SVD to calculate the standard t-score map in the template's space. The t-score for each patient $\mathbf{x}$ is calculated as:

$$\mathbf{t} = \frac{\mathbf{x} - \frac{1}{N_c}\Sigma_{i=1}^{N_c}\mathbf{c}_i}{\sqrt{\frac{N_c+1}{N_c(N_c+1)}\Sigma_{i=1}^{N_c}\left(\mathbf{c}_i - \frac{1}{N_c}\Sigma_{i=1}^{N_c}\mathbf{c}_i\right)^2}} \tag{7.7}$$

Here, $\mathbf{c}_i$ is the i-th control subject. Then the t-score maps are mapped to the individual native spaces using the transformations developed above. For each individual subject, we delineated the potential SVD lesions by applying the selected threshold to its t-score map and obtained the volume of the le-

sions. We then sorted the volumes of all subjects and chose the median as the threshold to distinguish normal and abnormal subjects in terms of SVD.

According to Table 7.1, our implementation of patch-based MIS-Boost outperforms the other two methods. It is clear that a simple method based on thresholds is unreliable since its sensitivity is low. This means it cannot detect abnormal subjects. Compared with the threshold-based method, the RF-based method improves the accuracy by 10%. In addition, the RF classifier achieves the trade-off between FPs and FNs since the gap between sensitivity and specificity is small. The use of voxels from the ROIs defined by the atlas enhances the accuracy by 5% compared to using voxels from the whole brain. Furthermore, our proposed model boosts the classification accuracy by an additional 5%. Apart from the high accuracy of classification, the sensitivity of MIS-Boost is high.

7.4 Discussion and Conclusion

We have presented a framework in which boosting-based MIL is used to learn patches for discrimination of normal or abnormal brain degeneration. To the best of our knowledge, this is the first work to automatically identify the SVD categories on CT.

A key feature of the proposed method is that individual CT images are formulated as bags, which enables the SVD identification without fine-grained annotations. We propose to use an atlas of SVD lesions derived from MR images. Compared with the low resolution of CT images, MR images are able to show brain lesions in detail. MR imaging is therefore regarded as the gold standard in the assessment of SVD. This provides prior knowledge where lesions occur frequently in the brain. In addition, our method has been verified on a large clinical CT dataset, which shows potentials of clinical use.

We have also shown that the classification results obtained using classic techniques such as RF are not as good as those achieved using our proposed approach. The proposed method also showed its strength compared to standard clinical approaches, where basic statistical features are used. Since CT images show a low signal-to-noise ratio, small lesions like SVD are difficult to be identified at a voxel. In contrast, patch-based features decrease the effect of noise.

In the future, the proposed method will be applied to a larger dataset including data from different clinical centres so that the framework can be tested more widely in terms of robustness and accuracy. More importantly, our final goal is to predict the outcome of stroke – whether the stroke patients will hemorrhage or not. This will help to reduce the rate of ICH significantly, which will improve quality of patients' lives and reduce the pressure for the public health services.

Chapter 8

Conclusion

8.1 Summary

The thesis has focused on addressing key challenges in medical image analysis. The methods proposed in this thesis make use of different machine learning techniques, including random forests, deep neural networks, and multi-instance learning.

The first contribution is an automated framework based on multi-scale patches and random forests for segmentation on poor quality medical images. In Chapter 3, this framework has been applied to segment WML on clinical CT images. Image patches across multiple scales provide rich context information. Using this information, lesions can be accurately segmented even when the image quality is poor. An existing WML atlas was used to guide the segmentation, which improves the efficiency of the framework. A comprehensive evaluation of the framework was performed based on large clinical trial datasets. The results obtained using the proposed framework were compared to those achieved by human experts, in terms of segmentation similarity, predicted lesion volume, and disease ratings using two scoring systems. Our results showed that the proposed method was comparable to human experts, which suggests our method is robust and applicable to clinical practice.

In medical image analysis, segmentation is a common challenge in many clinical applications. A generic segmentation method is useful as it can be used in different applications. With this in mind,

we have proposed a novel CNN architecture, namely DRINet, which was based on the DenseNet [11] and the Inception Net [15]. The three key components of the DRINet, which are the dense connection block, the residual inception block, and the unpooling block, widen and deepen the network over the classic FCN while the parameter space of the whole network can be controlled. In Chapter 4 The DRINet architecture has been validated on three different types of segmentation problems: Multi-class segmentation of brain CSF in CT images, multi-organ segmentation of abdomen in CT images, and multi-class segmentation of brain tumours in MR images. Compared to the classic approaches such as FCN and the U-Net [6], significant improvements were observed for the DRINet approach in all three experiments.

Although the DRINet solves general segmentation problems well, a more sophisticated framework is required for complex segmentation problems. More precisely, if the target of interest is highly variable in terms of position, size, and shape, and artefacts are present, a framework based on multiple CNNs is desired. In Chapter 5, we have proposed such a framework to segment the acute ischemic lesions in DW images. A particular challenge is that acute ischemic lesions can occur anywhere in the brain. Furthermore, the difference between real lesions and the artefacts is very subtle. As such, a single CNN cannot address the segmentation problem well due to FPs and FNs. To address this, we proposed the EDD Net which is an ensemble of two DeconvNets [2] and reduces the number of FPs and FNs. The second CNN, which is the MUSCLE Net evaluated the detections of the EDD Net and reduced the number of FPs further. Compared to the single CNN, we observed significant improvements achieved by the EDD Net and the MUSCLE Net, respectively.

In medical imaging, there are often only very few images which have been annotated. If models were developed on these limited images with annotations, the large amount of images without annotations remain unused. Using the images without annotations is likely to help to improve the performance of machine learning models. We have proposed a novel strategy to make use of the unannotated images to improve the model performance. The proposed approach was based on the self-supervised learning; the basic idea was training an extra CNN, which learns the image semantics using synthetic labels. Afterwards the learned weights were used to initialize the main CNN, which was trained on the annotated images. For medical images are typically static images in gray scale, existing applicable self-supervised methods, including patch relative positions and local context prediction, are likely

to learn trivial features. We have proposed to randomly disrupt the image context and used a CNN to learn to restore it. This forces the CNN to learn the image semantics. In Chapter 6, we have evaluated the proposed strategy in three different types of medical image analysis problems, namely classification, localization, and segmentation. These applications were based on different types of medical images, namely fetal ultrasound images, abdominal CT images, and brain MR images. The results demonstrated that 1) the proposed context restoration strategy improves CNN performance in all cases; 2) the context restoration learning provides more useful image semantics than existing methods, which lead to the best results; 3) in some cases, the performance achieved via self-supervised learning is comparable to full supervised learning.

Since annotating medical images is expensive in terms of time and human resource, there are several scenarios in which images can only be weakly annotated. An example is that only coarse-grained labels are available. In Chapter 7, we have proposed a framework based on multi-instance learning to address such problem. In the clinic, the SVD is usually assessed via CT or MR images and grading is performed according to well-established scoring systems. Usually, no fine-grained annotations of the lesions in medical images is performed. In the proposed approach, patches associated with the SVD lesions were extracted from individual CT images. The patch extraction was guided by an MRI-derived atlas, which defines the ROIs of SVD lesions. Each CT image was regarded as a bag comprising a number of patches (instances). The MIS-Boost [93] algorithm was adapted and applied on these bags. The learning results were a few distinctive instances discriminating different bags. The proposed framework was observed to perform better than classic machine learning methods such as the random forests, as well as existing clinical methods based on statistical scores and thresholds.

8.2 Limitations

Although the thesis has contributed to several key challenges in medical image analysis problems, these contributions still have some limitations.

Chapters 3, 4, and 5 have focused on segmentation problems and achieve promising results. However, given a set of images with manual segmentations, the upper bound accuracy that an automatic method

can achieve is still unknown. Empirically, a generic approach such as the DRINet is applied to achieve a baseline segmentation. Then techniques including adjusting the network parameter space are used to improve the baseline. However, it is difficult to understand if the developed model can be improved further and how much improvement can be expected.

Second, we have contributed a novel self-supervised learning strategy which restores image context, in order to make use of images without annotations. Our results showed that the CNNs solving different types of problems benefit from the self-supervised proxy training, which improves the CNN performance. We also observed that the performance improved more if the labelled training data was less. Furthermore, the context restoration was superior to existing self-supervised learning methods in terms of learning image semantics. However, these observations are only based on empirical evidence. It remains unknown whether there is a generally optimal self-supervised learning strategy or what is the best self-supervised learning strategy for individual tasks. It is also open how much improvement can be expected with certain amount of annotated data using a self-supervised learning strategy.

Finally, the analysis in Chapter 3 included annotations from multiple experts. The annotations, including pixel-level delineations and subject-level ratings, contain differences among experts. However, the model training in the work of all chapters has been based on single expert's delineations. These delineations are regarded as ground truth labels for training, which is likely to lead to a bias in the evaluation. As such, the models do not work well in some cases. For instance, the random forests model developed in Chapter 3 was observed to be more conservative than other experts, in terms of SVD lesion estimation.

8.3 Future Work

In the future, it would be very interesting to explore the following areas in more detail:

Theoretical Analysis: In medical image analysis, many problems can be well addressed using DNNs, such as image classification and segmentation. The most promising results are achieved using large amounts of annotated images. However, in medical imaging it is difficult to build datasets as large as

the Microsoft COCO [233]. Therefore, one has to be careful with regards to underfitting and overfitting. To address underfitting problems, a theoretical analysis is necessary to identify the performance upper bound and how one can reach the upper bound. To address overfitting problems, learning to generalize from images with limited annotations, weak annotations or without annotations is the key. This requires solid theoretical analysis which does not exist.

Data Auditing: In an image dataset comprised of natural images, objects of interest (e.g. cars, houses) can be accurately recognized by most human observers. However, given a medical image, it is likely that different experts may have different opinions with regards to the annotation. Ideally each image should be annotated by a number of experts and a consensus can be computed and used for model development. However, this is too expensive in terms of time and human resources. In practice, a dataset is usually annotated by a small group of experts. Each of the experts annotate a subset of images and each image is annotated by no more than a few experts. Before model development, the quality of annotations by different experts should be evaluated so that a reliable consensus can be achieved.

Extracting Clinically Useful Information: Medical image analysis provides clinicians with useful tools for quantification, diagnosis and treatment planning. For classification tasks, the output is a predicted categorical label. For detection tasks, the output is usually a bounding box highlighting objects of interest. For segmentation tasks, the output is typically a soft probability map, which can be converted to binary maps. Then the volume of segmentation can be quantified. The binary segmentation map can be used for geometric reconstruction and visualization. In addition, more clinically useful information can be explored, such as the correlation between the measured quantities and disease outcomes. For instance, we have contributed to analysing medical images from stroke patients and we have achieved good results on SVD segmentation, brain CSF segmentation and acute ischemic lesion segmentation. Afterwards, it is clinically useful to study if measurements on these segmentations relate to the ICH. Furthermore, follow-up scans can be used to assess lesion evolvement. Understanding the lesion evolvement helps clinicians improve their diagnosis and treatment.

Bibliography

[1] C. Farrell, F. Chappell, P. Armitage, P. Keston, A. MacLullich, S. Shenkin, and J. Wardlaw, "Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years," *European Radiology*, vol. 19, no. 1, pp. 177–183, 2009.

[2] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528, 2015.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems*, pp. 153–160, 2007.

[4] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

[5] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[11] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2016.

[12] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, pp. 3859–3869, 2017.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2018.

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on Artificial Intelligence*, pp. 4278–4284, 2017.

[16] J. Beutel, H. L. Kundel, and R. L. Van Metter, *Handbook of medical imaging: physics and psychophysics*, vol. 1. Spie Press, 2000.

[17] M. Yaffe and J. Rowlands, "X-ray detectors for digital radiography," *Physics in Medicine & Biology*, vol. 42, no. 1, p. 1, 1997.

[18] J. Hsieh *et al.*, *Computed tomography: principles, design, artifacts, and recent advances*. SPIE, 2009.

[19] P. A. Rinck, *Magnetic resonance in medicine a critical introduction*. BoD, 2018.

[20] P. R. Hoskins, K. Martin, and A. Thrush, *Diagnostic ultrasound: physics and equipment*. Cambridge University Press, 2010.

[21] V. I. Mikla and V. V. Mikla, *Medical imaging technology*. Elsevier, 2013.

[22] L.-O. Wahlund, F. Barkhof, F. Fazekas, L. Bronge, M. Augustin, M. Sjögren, A. Wallin, H. Adèr, D. Leys, L. Pantoni, *et al.*, "A new rating scale for age-related white matter changes applicable to MRI and CT," *Stroke*, vol. 32, no. 6, pp. 1318–1322, 2001.

[23] J. Van Swieten, A. Hijdra, P. Koudstaal, and J. Van Gijn, "Grading white matter lesions on CT and MRI: a simple scale," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 53, no. 12, pp. 1080–1083, 1990.

[24] L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage: Clinical*, 2017.

[25] L. Chen, A. Carlton Jones, G. Mair, R. Patel, A. Gontsarova, J. Ganesalingam, N. Math, A. Dawson, A. Basaam, D. Cohen, *et al.*, "Rapid automated quantification of cerebral leukoaraiosis on CT: a multicentre validation study," *Radiology*, 2018.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[27] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, p. 170117, 2017.

[28] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T O'Brien, F. Barkhof, O. R. Benavente, *et al.*, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *The Lancet Neurology*, vol. 12, no. 8, pp. 822–838, 2013.

[29] N. R. Sims and H. Muyderman, "Mitochondria, oxidative metabolism and cell death in stroke," *Biochimica et biophysica acta (BBA)- molecular basis of disease*, vol. 1802, no. 1, pp. 80–91, 2010.

[30] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[31] G. Donnan, M. Fisher, M. Macleod, and S. M. Davis, "Stroke," *The Lancet*, vol. 371, pp. 1612–1623, 2008.

[32] A. Durukan and T. Tatlisumak, "Acute ischemic stroke: overview of major experimental rodent models, pathophysiology, and therapy of focal cerebral ischemia," *Pharmacology Biochemistry and Behavior*, vol. 87, pp. 179–197, 2007.

[33] U. Dirnagl, C. Iadecola, and M. A. Moskowitz, "Pathobiology of ischaemic stroke: an integrated view," *Trends in Neurosciences*, vol. 22, no. 9, pp. 391–397, 1999.

[34] J. M. Wardlaw, V. Murray, E. Berge, G. Del Zoppo, P. Sandercock, R. L. Lindley, and G. Cohen, "Recombinant tissue plasminogen activator for acute ischaemic stroke: An updated systematic review and meta-analysis," *The Lancet*, vol. 379, pp. 2364–2372, 2012.

[35] M. Wintermark, G. W. Albers, A. V. Alexandrov, J. R. Alger, R. Bammer, J.-C. Baron, S. Davis, B. M. Demaerschalk, C. P. Derdeyn, G. A. Donnan, *et al.*, "Acute stroke imaging research roadmap," *American Journal of Neuroradiology*, vol. 29, no. 5, pp. e23–e30, 2008.

[36] M. M. a. Conijn, R. P. Kloppenborg, A. Algra, W. P. T. M. Mali, L. J. Kappelle, K. L. Vincken, Y. Van Der Graaf, and M. I. Geerlings, "Cerebral small vessel disease and risk of death, ischemic stroke, and cardiac complications in patients with atherosclerotic disease: The second

manifestations of arterial disease-magnetic resonance (SMART-MR) study," *Stroke*, vol. 42, pp. 3105–3109, 2011.

[37] L. Pantoni, "Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges," *The Lancet Neurology*, vol. 9, no. 7, pp. 689–701, 2010.

[38] M. A. Williams and N. R. Relkin, "Diagnosis and management of idiopathic normal-pressure hydrocephalus," *Neurology: Clinical Practice*, vol. 3, no. 5, pp. 375–385, 2013.

[39] I.-. C. Group *et al.*, "Association between brain imaging signs, early and late outcomes, and response to intravenous alteplase after acute ischaemic stroke in the third International Stroke Trial (IST-3): secondary analysis of a randomised controlled trial," *The Lancet Neurology*, vol. 14, no. 5, pp. 485–496, 2015.

[40] P. Fotiadis, S. van Rooden, J. van der Grond, A. Schultz, S. Martinez-Ramirez, E. Auriel, Y. Reijmer, A. M. van Opstal, A. Ayres, K. M. Schwab, *et al.*, "Cortical atrophy in patients with cerebral amyloid angiopathy: a case-control study," *The Lancet Neurology*, vol. 15, no. 8, pp. 811–819, 2016.

[41] C. M. Dunham, D. A. Hoffman, G. S. Huang, L. A. Omert, D. J. Gemmel, and R. Merrell, "Traumatic intracranial hemorrhage correlates with preinjury brain atrophy, but not with antithrombotic agent use: a retrospective study," *PloS one*, vol. 9, no. 10, p. e109473, 2014.

[42] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997.

[43] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[44] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] M. P. Perrone and L. N. Cooper, "When networks disagree: ensemble methods for hybrid neural networks," in *Artificial Neural Networks for Speech and Vision*, pp. 126–142, 1993.

[46] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[47] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999.

[48] B. Clarke, "Comparing bayes model averaging and stacking when model approximation error cannot be ignored," *Journal of Machine Learning Research*, vol. 4, pp. 683–712, 2003.

[49] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[50] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*, pp. 195–201, 1995.

[51] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, p. 386, 1958.

[52] P. Werbos, *Beyond regression: new tools for prediction and analysis in the behavior science*. PhD thesis, Harvard University, 1974.

[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California University San Diego, 1985.

[54] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[55] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[56] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[57] K. Fukushima, "Cognitron: a self-organizing multilayered neural network," *Biological Cybernetics*, vol. 20, no. 3-4, pp. 121–136, 1975.

[58] K. Fukushima and S. Miyake, "Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and Cooperation in Neural Nets*, pp. 267–285, Springer, 1982.

[59] J. Bouvrie, "Notes on convolutional neural networks," tech. rep., Massachusetts Institute of Technology, 2006.

[60] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning*, pp. 807–814, 2010.

[61] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

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

[65] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1529–1537, 2015.

[66] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[67] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[68] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: learning dense volumetric segmentation from sparse annotation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016.

[69] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, 2018.

[70] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1175–1183, 2017.

[71] A. Krishnakumar, "Active learning literature survey," tech. rep., University of California, Santa Cruz, 2007.

[72] X. Zhu, "Semi-supervised learning literature survey," tech. rep., University of Wisconsin, Madison, 2005.

[73] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2006.

[74] D. J. Miller and H. S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," in *Advances in Neural Information Processing Systems*, pp. 571–577, 1997.

[75] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[76] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the International Conference on Machine Learning*, pp. 19–26, 2001.

[77] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems*, pp. 321–328, 2004.

[78] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 57–64, 2005.

[79] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the International Conference on Machine Learning*, vol. 99, pp. 200–209, 1999.

[80] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou, "Convex and scalable weakly labeled SVMs," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2151–2188, 2013.

[81] Y.-F. Li and Z.-H. Zhou, "Towards making unlabeled data never hurt," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 175–188, 2015.

[82] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Conference on Computational Learning Theory*, pp. 92–100, 1998.

[83] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.

[84] Z.-H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.

[85] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

[86] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proceedings of the European Conference on Computer Vision*, pp. 649–666, 2016.

[87] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proceedings of the European Conference on Computer Vision*, pp. 69–84, 2016.

[88] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 37–45, 2015.

[89] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2701–2710, 2017.

[90] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proceedings of the International Conference on Machine Learning*, pp. 737–744, 2009.

[91] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[92] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, pp. 577–584, 2003.

[93] E. Akbas, B. Ghanem, and N. Ahuja, "MIS-Boost: multiple instance selection boosting," *arXiv preprint arXiv:1109.2388*, 2011.

[94] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance learning with randomized trees," in *Proceedings of the European Conference on Computer Vision*, pp. 29–42, 2010.

[95] Z.-H. Zhou and M.-L. Zhang, "Solving multi-instance problems with classifier ensemble based on constructive clustering," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 155–170, 2007.

[96] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, A. D. N. Initiative, *et al.*, "Multiple instance learning for classification of dementia in brain MRI," *Medical Image Analysis*, vol. 18, no. 5, pp. 808–818, 2014.

[97] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[98] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Transactions on Multimedia*, vol. 12, no. 2, pp. 131–141, 2010.

[99] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 131–167, 1999.

[100] D. C. Brabham, "Crowdsourcing as a model for problem solving: an introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[101] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC Press, 2012.

[102] R. Urner, S. B. David, and O. Shamir, "Learning from weak teachers," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1252–1260, 2012.

[103] R. Rossi, C. Joachim, C. Geroldi, M. M. Esiri, A. D. Smith, and G. B. Frisoni, "Pathological validation of a CT-based scale for subcortical vascular disease," *Dementia and geriatric cognitive disorders*, vol. 19, no. 2-3, pp. 61–66, 2005.

[104] N. Sanossian, K. A. Fu, D. S. Liebeskind, S. Starkman, S. Hamilton, J. P. Villablanca, A. M. Burgos, R. Conwit, and J. L. Saver, "Utilization of emergent neuroimaging for thrombolysis-eligible stroke patients," *Journal of Neuroimaging*, vol. 27, no. 1, pp. 59–64, 2017.

[105] W.-S. Ryu, S.-H. Woo, D. Schellingerhout, M. U. Jang, K.-J. Park, K.-S. Hong, S.-W. Jeong, J.-Y. Na, K.-H. Cho, J.-T. Kim, *et al.*, "Stroke outcomes are worse with larger leukoaraiosis volumes," *Brain*, vol. 140, no. 1, pp. 158–170, 2016.

[106] N. Henninger, S. Izzy, R. Carandang, W. Hall, and S. Muehlschlegel, "Severe leukoaraiosis portends a poor outcome after traumatic brain injury," *Neurocritical Care*, vol. 21, no. 3, pp. 483–495, 2014.

[107] A. Charidimou, M. Pasi, M. Fiorelli, S. Shams, R. von Kummer, L. Pantoni, and N. Rost, "Leukoaraiosis, cerebral hemorrhage, and outcome after intravenous thrombolysis for acute ischemic stroke," *Stroke*, vol. 47, no. 9, pp. 2364–2372, 2016.

[108] L. Willer, I. Havsteen, C. Ovesen, A. F. Christensen, and H. Christensen, "Computed tomography–verified leukoaraiosis is a risk factor for post-thrombolytic hemorrhage," *Journal of Stroke and Cerebrovascular Diseases*, vol. 24, no. 6, pp. 1126–1130, 2015.

[109] M. Alachkar, "Neuroimaging in dementia: how best to use the guidelines?," *Psychiatric Bulletin*, vol. 38, no. 3, pp. 137–138, 2014.

[110] T. Kuruvilla, R. Zheng, B. Soden, S. Greef, and I. Lyburn, "Neuroimaging in a memory assessment service: a completed audit cycle," *Psychiatric Bulletin*, vol. 38, no. 1, pp. 24–28, 2014.

[111] R. Riello, C. Albini, S. Galluzzi, P. Pasqualetti, and G. Frisoni, "Prescription practices of diagnostic imaging in dementia: a survey of 47 Alzheimer's Centres in Northern Italy," *International Journal of Geriatric Psychiatry*, vol. 18, no. 7, pp. 577–585, 2003.

[112] M. Simoni, L. Li, N. L. Paul, B. E. Gruter, U. G. Schulz, W. Küker, and P. M. Rothwell, "Age- and sex-specific rates of leukoaraiosis in TIA and stroke patients Population-based study," *Neurology*, vol. 79, no. 12, pp. 1215–1222, 2012.

[113] P. Scheltens, T. Erkinjunti, D. Leys, L.-O. Wahlund, D. Inzitari, T. del Ser, F. Pasquier, F. Barkhof, R. Mäntylä, J. Bowler, *et al.*, "White matter changes on CT and MRI: an overview of visual rating scales," *European Neurology*, vol. 39, no. 2, pp. 80–89, 1998.

[114] L. Pantoni, M. Simoni, G. Pracucci, R. Schmidt, F. Barkhof, D. Inzitari, *et al.*, "Visual rating scales for age-related white matter changes (leukoaraiosis)," *Stroke*, vol. 33, no. 12, pp. 2827–2833, 2002.

[115] P. Sandercock, J. M. Wardlaw, R. I. Lindley, M. Dennis, G. Cohen, G. Murray, K. Innes, G. Venables, A. Czlonkowska, A. Kobayashi, *et al.*, "The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial," *Lancet*, vol. 379, no. 9834, pp. 2352–2363, 2012.

[116] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[117] L. Chen, T. Tong, C. P. Ho, R. Patel, D. Cohen, A. C. Dawson, O. Halse, O. Geraghty, P. E. Rinne, C. J. White, *et al.*, "Identification of cerebral small vessel disease using multiple instance learning," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 523–530, 2015.

[118] C. Ledig, W. Shi, W. Bai, and D. Rueckert, "Patch-based evaluation of image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3065–3072, 2014.

[119] C. Ledig, R. A. Heckemann, A. Hammers, J. C. Lopez, V. F. Newcombe, A. Makropoulos, J. Lötjönen, D. K. Menon, and D. Rueckert, "Robust whole-brain segmentation: application to traumatic brain injury," *Medical Image Analysis*, vol. 21, no. 1, pp. 40–58, 2015.

[120] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Wiley StatsRef: Statistics Reference Online*, 2006.

[121] S. Vanbelle and A. Albert, "A bootstrap method for comparing correlated kappa coefficients," *Journal of Statistical Computation and Simulation*, vol. 78, no. 11, pp. 1009–1015, 2008.

[122] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[123] C. R. Gillebert, G. W. Humphreys, and D. Mantini, "Automated delineation of stroke lesions using brain CT images," *NeuroImage: Clinical*, vol. 4, pp. 540–548, 2014.

[124] C. Herweh, P. A. Ringleb, G. Rauch, S. Gerry, L. Behrens, M. Möhlenbruch, R. Gottorf, D. Richter, S. Schieber, and S. Nagel, "Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients," *International Journal of Stroke*, vol. 11, no. 4, pp. 438–445, 2016.

[125] P. Bentley, J. Ganesalingam, A. L. C. Jones, K. Mahady, S. Epton, P. Rinne, P. Sharma, O. Halse, A. Mehta, and D. Rueckert, "Prediction of stroke thrombolysis outcome using CT brain machine learning," *NeuroImage: Clinical*, vol. 4, pp. 635–640, 2014.

[126] W. N. Whiteley, K. B. Slot, P. Fernandes, P. Sandercock, and J. Wardlaw, "Risk factors for intracranial hemorrhage in acute ischemic stroke patients treated with recombinant tissue plasminogen activator a systematic review and meta-analysis of 55 studies," *Stroke*, vol. 43, no. 11, pp. 2904–2909, 2012.

[127] E. Smith, J. Rosand, K. Knudsen, E. Hylek, and S. Greenberg, "Leukoaraiosis is associated with warfarin-related hemorrhage following ischemic stroke," *Neurology*, vol. 59, no. 2, pp. 193–197, 2002.

[128] M. Lou, A. Al-Hazzani, R. P. Goddeau, V. Novak, and M. Selim, "Relationship between white-matter hyperintensities and hematoma volume and growth in patients with intracerebral hemorrhage," *Stroke*, vol. 41, no. 1, pp. 34–40, 2010.

[129] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[130] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," in *Proceedings of the International Conference on Learning Representations*, 2018.

[131] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[132] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning*, pp. 448–456, 2015.

[133] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[134] N. Sanossian, K. A. Fu, D. S. Liebeskind, S. Starkman, S. Hamilton, J. P. Villablanca, A. M. Burgos, R. Conwit, and J. L. Saver, "Utilization of emergent neuroimaging for thrombolysis-eligible stroke patients," *Journal of Neuroimaging*, vol. 27, no. 1, pp. 59–64, 2017.

[135] I. K. Pople, "Hydrocephalus and shunts: what the neurologist should know," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. suppl 1, pp. i17–i22, 2002.

[136] A. V. Kulkarni, J. M. Drake, D. C. Armstrong, and P. B. Dirks, "Measurement of ventricular size: reliability of the frontal and occipital horn ratio compared to subjective assessment," *Pediatric neurosurgery*, vol. 31, no. 2, pp. 65–70, 1999.

[137] F. Pasquier, D. Leys, J. G. Weerts, F. Mounier-Vehier, F. Barkhof, and P. Scheltens, "Inter-and intraobserver reproducibility of cerebral atrophy assessment on mri scans with hemispheric infarcts," *European neurology*, vol. 36, no. 5, pp. 268–272, 1996.

[138] T. Sandor, D. Metcalf, and Y.-J. Kim, "Segmentation of brain CT images using the concept of region growing," *International Journal of Bio-Medical Computing*, vol. 29, no. 2, pp. 133–147, 1991.

[139] U. E. Ruttimann, E. M. Joyce, D. E. Rio, and M. J. Eckardt, "Fully automated segmentation of cerebrospinal fluid in computed tomography," *Psychiatry Research: Neuroimaging*, vol. 50, no. 2, pp. 101–119, 1993.

[140] T. H. Lee, M. F. A. Fauzi, and R. Komiya, "Segmentation of CT brain images using K-means and EM clustering," in *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*, pp. 339–344, 2008.

[141] T. H. Lee, M. F. A. Fauzi, and R. Komiya, "Segmentation of CT brain images using unsupervised clusterings," *Journal of Visualization*, vol. 12, no. 2, pp. 131–138, 2009.

[142] W. Chen and K. Najarian, "Segmentation of ventricles in brain CT images using gaussian mixture model method," in *Proceedings of the International Conference on Complex Medical Engineering*, pp. 1–6, 2009.

[143] V. Gupta, W. Ambrosius, G. Qian, A. Blazejewska, R. Kazmierski, A. Urbanik, and W. L. Nowinski, "Automatic segmentation of cerebrospinal fluid, white and gray matter in unenhanced computed tomography images," *Academic Radiology*, vol. 17, no. 11, pp. 1350–1358, 2010.

[144] L. Poh, V. Gupta, A. Johnson, R. Kazmierski, and W. L. Nowinski, "Automatic segmentation of ventricular cerebrospinal fluid from ischemic stroke CT images," *Neuroinformatics*, vol. 10, no. 2, pp. 159–172, 2012.

[145] X. Qian, J. Wang, S. Guo, and Q. Li, "An active contour model for medical image segmentation with application to brain CT image," *Medical Physics*, vol. 40, no. 2, 2013.

[146] X. Qian, Y. Lin, Y. Zhao, X. Yue, B. Lu, and J. Wang, "Objective ventricle segmentation in brain CT with ischemic stroke based on anatomical knowledge," *BioMed Research International*, vol. 2017, 2017.

[147] M. G. Linguraru, J. A. Pura, V. Pamulapati, and R. M. Summers, "Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT," *Medical Image Analysis*, vol. 16, no. 4, pp. 904–914, 2012.

[148] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from ct volumes," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 149–157, 2016.

[149] K. Karasawa, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, C. Chu, G. Zheng, D. Rueckert, and K. Mori, "Multi-atlas pancreas segmentation: atlas selection based on vessel structure," *Medical Image Analysis*, vol. 39, pp. 18–28, 2017.

[150] T. Okada, R. Shimada, M. Hori, M. Nakamoto, Y.-W. Chen, H. Nakamura, and Y. Sato, "Automated segmentation of the liver from 3D CT images using probabilistic atlas and multilevel statistical shape model," *Academic Radiology*, vol. 15, no. 11, pp. 1390–1403, 2008.

[151] Z. Wang, K. K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert, "Geodesic patch-based segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 666–673, 2014.

[152] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.

[153] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal, and D. Rueckert, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Medical Image Analysis*, vol. 23, no. 1, pp. 92–104, 2015.

[154] C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, and K. Mori, "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 165–172, 2013.

[155] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *arXiv preprint arXiv:1702.00045*, 2017.

[156] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, "Hierarchical 3D fully convolutional networks for multi-organ segmentation," *arXiv preprint arXiv:1704.06382*, 2017.

[157] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," *arXiv preprint arXiv:1707.04912*, 2017.

[158] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, 2018.

[159] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita, "Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting," in *Deep Learning and Data Labeling for Medical Applications*, pp. 111–120, Springer, 2016.

[160] P. Hu, F. Wu, J. Peng, Y. Bao, F. Chen, and D. Kong, "Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets," *International journal of computer assisted radiology and surgery*, vol. 12, no. 3, pp. 399–411, 2017.

[161] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection," *The Cancer Imaging Archive*, 2017.

[162] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[163] A. Wouters, P. Dupont, B. Norrving, R. Laage, G. Thomalla, G. W. Albers, V. Thijs, and R. Lemmens, "Prediction of stroke onset is improved by relative fluid-attenuated inversion recovery and perfusion imaging compared to the visual diffusion-weighted imaging/fluid-attenuated inversion recovery mismatch," *Stroke*, vol. 47, no. 10, pp. 2559–2564, 2016.

[164] P. Rinne, M. Hassan, D. Goniotakis, K. Chohan, P. Sharma, D. Langdon, D. Soto, and P. Bentley, "Triple dissociation of attention networks in stroke according to lesion location," *Neurology*, vol. 81, no. 9, pp. 812–820, 2013.

[165] M. G. Dwyer, N. Bergsland, E. Saluste, J. Sharma, Z. Jaisani, J. Durfee, N. Abdelrahman, A. Minagar, R. Hoque, F. E. Munschauer, *et al.*, "Application of hidden markov random field approach for quantification of perfusion/diffusion mismatch in acute ischemic stroke," *Neurological Research*, vol. 30, no. 8, pp. 827–834, 2008.

[166] A. L. Martel, S. J. Allder, G. S. Delay, P. S. Morgan, and A. R. Moody, "Measurement of infarct volume in stroke patients using adaptive segmentation of diffusion weighted MR images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 22–31, 1999.

[167] W. Charoensuk, N. Covavisaruch, S. Lerdlum, and Y. Likitjaroen, "Acute stroke brain infarct segmentation in DWI images," *International Journal of Pharma Medicine and Biological Sciences*, vol. 4, no. 2, p. 115, 2015.

[168] M. A. Jacobs, R. A. Knight, H. Soltanian-Zadeh, Z. G. Zheng, A. V. Goussev, D. J. Peck, J. P. Windham, and M. Chopp, "Unsupervised segmentation of multiparameter MRI in experimental cerebral ischemia with comparison to T2, diffusion, and ADC MRI parameters and histopathological validation," *Journal of Magnetic Resonance Imaging*, vol. 11, no. 4, pp. 425–437, 2000.

[169] M. Li, L. Ai, H. He, Z. Zheng, B. Lv, W. Li, J. Yi, and X. Chen, "Segmentation of infarct in acute ischemic stroke from MR apparent diffusion coefficient and trace-weighted images," in *Proceedings of the International Symposium on Multispectral Image Processing and Pattern Recognition*, pp. 74971U–74971U, 2009.

[170] H. Soltanian-Zadeh, H. Bagher-Ebadian, J. R. Ewing, P. D. Mitsias, A. Kapke, M. Lu, Q. Jiang, S. C. Patel, and M. Chopp, "Multiparametric iterative self-organizing data analysis of ischemic lesions using pre-or post-Gd T1 MRI," *Cerebrovascular Diseases*, vol. 23, no. 2-3, pp. 91–102, 2006.

[171] Y.-H. Mah, R. Jager, C. Kennard, M. Husain, and P. Nachev, "A new method for automated high-dimensional lesion segmentation evaluated in vascular injury and applied to the human occipital lobe," *Cortex*, vol. 56, pp. 51–63, 2014.

[172] H. B. van der Worp and J. van Gijn, "Acute ischemic stroke," *New England Journal of Medicine*, vol. 357, no. 6, pp. 572–579, 2007.

[173] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, *et al.*, "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Medical Image Analysis*, vol. 35, pp. 250–269, 2017.

[174] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[175] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: a retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[176] P. D. Chang, "Fully convolutional neural networks with hyperlocal features for brain tumor segmentation," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 4–9, 2016.

[177] T. K. Lun and W. Hsu, "Brain tumor segmentation using deep convolutional neural network," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 26–29, 2016.

[178] R. randhawa, A. Modi, P. Jain, and P. Warier, "Improving segment boundary classification for brain tumor segmentation and longitudinal disease progression," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 53–56, 2016.

[179] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "DeepMedic on brain tumor segmentation," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 18–22, 2016.

[180] A. Ellwaa, A. Hussein, E. AlNaggar, M. Zidan, M. Zaki, M. A. Ismail, and N. M. Ghanem, "Brain tumor segmentation using random forest trained on iterative selected patients," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 14–17, 2016.

[181] L. Lefkovits, S. Lefkovits, and L. Szilágyi, "Brain tumor segmentation with optimized random forest," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 30–34, 2016.

[182] L. L. Folgoc, A. V. Nori, J. Alvarez-Valle, R. Lowe, and A. Criminisi, "Segmentation of brain tumors via cascades of lifted decision forests," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 35–39, 2016.

[183] B. Song, C.-R. Chou, A. Huang, and M.-C. Liu, "Anatomy-guided brain tumor segmentation and classification," in *Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection"*, pp. 61–64, 2016.

[184] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, "Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI," *Ischemic Stroke Lesion Segmentation*, p. 13, 2015.

[185] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[186] C. Feng, D. Zhao, and M. Huang, "Segmentation of stroke lesions in multi-spectral MR images using bias correction embedded FCM and three phase level set," *Ischemic Stroke Lesion Segmentation*, p. 3, 2015.

[187] H.-L. Halme, A. Korvenoja, and E. Salli, "ISLES (SISS) challenge 2015: segmentation of stroke lesions using spatial normalization, Random Forest classification and contextual clustering," in *Proceedings of the International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 211–221, 2015.

[188] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "DCAN: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, 2016.

[189] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 556–564, 2015.

[190] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[191] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, 2014.

[192] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[193] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O?Regan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 246–254, 2016.

[194] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[195] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[196] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017.

[197] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[198] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

[199] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016.

[200] D. Jayaraman and K. Grauman, "Slow and steady feature analysis: higher order temporal coherence in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3852–3861, 2016.

[201] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2015.

[202] J. Walker, A. Gupta, and M. Hebert, "Dense optical flow prediction from a static image," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2443–2451, 2015.

[203] S. Purushwalkam and A. Gupta, "Pose from action: Unsupervised learning of pose features based on motion," *arXiv preprint arXiv:1609.05420*, 2016.

[204] P. Sermanet, C. Lynch, J. Hsu, and S. Levine, "Time-contrastive networks: Self-supervised learning from multi-view observation," *arXiv preprint arXiv:1704.06888*, 2017.

[205] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *Proceedings of the European Conference on Computer Vision*, pp. 527–544, 2016.

[206] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5729–5738, 2017.

[207] D. Jayaraman and K. Grauman, "Learning image representations tied to ego-motion," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1413–1421, 2015.

[208] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proceedings of the European Conference on Computer Vision*, pp. 801–816, 2016.

[209] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proceedings of the British Machine Vision Conference*, pp. 1–11, 2017.

[210] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for Spinal MRIs," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 294–302, Springer, 2017.

[211] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[212] NHS Screening Programmes, *Fetal Anomaly Screen Programme Handbook*. NHS, 2015.

[213] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[214] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[215] B. de Vos, J. Wolterink, P. de Jong, T. Leiner, M. Viergever, and I. Isgum, "ConvNet-based localization of anatomical structures in 3D medical images," *IEEE Transactions on Medical Imaging*, 2017.

[216] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems*, pp. 109–117, 2011.

[217] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6874–6883, 2017.

[218] F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman, "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging," *American Journal of Neuroradiology*, vol. 8, no. 3, pp. 421–426, 1987.

[219] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[220] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, *et al.*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.

[221] M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot, "Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI," *Hippocampus*, vol. 19, no. 6, pp. 579–587, 2009.

[222] A. V. Dalca, R. Sridharan, L. Cloonan, K. M. Fitzpatrick, A. Kanakis, K. L. Furie, J. Rosand, O. Wu, M. Sabuncu, N. S. Rost, *et al.*, "Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 773–780, 2014.

[223] M. Chawla, S. Sharma, J. Sivaswamy, and L. Kishore, "A method for automatic detection and classification of stroke from brain CT images," in *Proceedings of the IEEE Conference on Engineering in Medicine and Biology Society*, pp. 3581–3584, 2009.

[224] N. Takahashi, D.-Y. Tsai, Y. Lee, T. Kinoshita, and K. Ishii, "Z-score mapping method for extracting hypoattenuation areas of hyperacute stroke in unenhanced CT," *Academic Radiology*, vol. 17, no. 1, pp. 84–92, 2010.

[225] C. R. Gillebert, G. W. Humphreys, and D. Mantini, "Automated delineation of stroke lesions using brain CT images," *NeuroImage: Clinical*, vol. 4, pp. 540–548, 2014.

[226] Z.-h. Zhou, "Multi-instance learning: a survey," tech. rep., National Laboratory for Novel Software Technology, Nanjing, 2004.

[227] C. Leistner, A. Saffari, and H. Bischof, "MIForests: multiple-instance learning with randomized trees," in *Proceedings of the European Conference on Computer Vision*, pp. 29–42, 2010.

[228] Q. Zhang and S. Goldman, "EM-DD: an improved multiple-instance learning technique," in *Advances in Neural Information Processing Systems*, pp. 1073–1080, 2001.

[229] C. Rorden, L. Bonilha, J. Fridriksson, B. Bender, and H. O. Karnath, "Age-specific CT and MRI templates for spatial normalization," *NeuroImage*, vol. 61, pp. 957–965, 2012.

[230] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[231] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, "Approximate solution of the trust region problem by minimization over two-dimensional subspaces," *Mathematical Programming*, vol. 40, no. 1-3, pp. 247–263, 1988.

[232] J. Ramírez, J. Górriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, M. López, I. Álvarez, and P. Padilla, "Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification," *Neuroscience Letters*, vol. 472, no. 2, pp. 99–103, 2010.

[233] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.