Deep Learning in Medical Imaging: A Review

Upul Senanayake[1] Jian Kang[1] Matthew Gibson[1]
Arathy Satheesh Babu[1] Anastasia Levenkova[1] Gihan Samarasinghe[1]
Arcot Sowmya[1]

[1] University of New South Wales, Australia
{upul.senanayake,jian.kang,matthew.gibson1}@unsw.edu.au
{a.satheeshbabu,a.levenkova,gihan.samarasinghe}@unsw.edu.au
a.sowmya@unsw.edu.au

Technical Report
UNSW-CSE-TR-201618
December 2016

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Computer vision and machine learning techniques have usually found their way into medical image analysis as there are strong synergies between the two fields. Medical imaging professionals typically draw inspiration from computer vision techniques to register images from multiple modalities, segment the regions of interest (ROI) and measure ROIs to streamline the analysis pipeline, and from machine learning techniques to build models that are capable of integral tasks such as anomaly detection, progression tracking and recognition. Historically, we have seen fast adaptation of computer vision techniques and with the recent significant improvements demonstrated by the use of deep learning techniques in other fields, we believe that the medical imaging professional can benefit from it as well. Deep learning is a collection of techniques that consists of new as well as old algorithms from the neural networks community. Ideas such as convolutional neural networks and autoencoders go back to the 1980s and 1990s, while concepts such as deep belief networks and long short term memory in recurrent neural networks are relatively new. Since its advent, deep learning has been mostly perceived as a black-box and its adaptation to medical imaging has been relatively slow. In this report, we attempt to untangle the techniques collectively known as deep learning and present their building blocks concisely. We also present a comprehensive review of the state-of-the-art applications of deep learning techniques in medical imaging. It is our belief that we can inspire more medical professionals to adapt deep learning techniques to their analysis pipeline and further refine the techniques that have surfaced so far.

1 Introduction

Advances in computer vision (CV) and artificial intelligence (AI) have always played a large role in medical image analysis, where the specific techniques used are often inspired by their counterparts in CV and AI. There are a number of medical imaging techniques that are used effectively today to diagnose various diseases of humans as well as animals where applicable [1]. Some of the well known medical imaging modalities include X-ray, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). The inherent complexity of medical images has garnered a lot of interest from computer scientists to adapt CV techniques to medical image analysis [2]. Typically in a medical image analysis scenario, the process starts by registering images from multiple modalities, then they are segmented to find regions of interest (ROI). Once ROIs are found, different types of analysis can be carried out, including anomaly detection, measurement of specific ROI parameters, progression tracking and recognition. Any of these techniques in the analysis pipeline relies heavily on the data representation, which is better known as features [3]. Conventional wisdom is that the best features are those hand designed by engineers and domain experts to suit a specific application, however general CV features like SIFT features have been useful as well [4]. In general, it takes a lot of effort to find relevant and efficient features and a lot more time to validate them for a specific problem. This has been the inspiration behind the application of deep learning approaches in medical image analysis, as deep learning promises to eliminate the feature design step entirely in most applications. In fact, deep learning has made major advances in many problems that have been challenging for the AI community [5]. A good description of such applications can be found [6]. In this technical report, we aim to provide an in-depth review of state-of-the-art deep learning techniques used in medical image analysis. We seek to pay particular attention to the effective application of computer vision techniques in medical imaging, as the inherent complexity of medical images presents a number of challenges. The advent of deep learning techniques was enabled mainly due to the advances in computational resources and the ubiquitous availability of annotated data. While medical imaging shares the advances in computational resources, availability of annotated data still presents a significant problem. In addition, medical images are typically gray-scale images whereas many CV techniques are typically developed for three channel images (R-Red, G-Green, B-Blue). Medical images are typically volumetric in nature as opposed to natural images which makes it more challenging to apply CV techniques directly. While volumetric convolution is a possibility, it is still very much an active research area. Conversely, medical images are inherently shift and translational invariant, which is not the case with natural images used to train CV techniques. Another perceived advantage of medical images compared to natural images is that they often come in multiple modalities which also presents the challenge of fusing information from disparate domains including multiple images and textual information. Due to these inherent differences, we observe that deep learning techniques as used in CV will have to be amended before application to biomedical images. In this report, we attempt to review what has been achieved so far by presenting a comprehensive review of the deep learning applications in medical imaging.

To the best of our knowledge, this is the first such review, although some re-

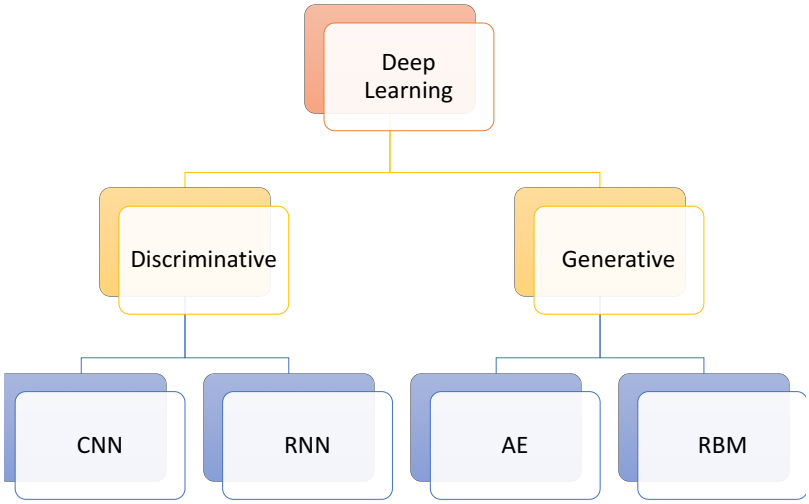

Figure 1.1: Overview of Deep Learning Methods

views on deep learning in bioinformatics research can be found [7, 8]. Greenspan et al. [9] provided a guest editorial that covered some applications of deep learning in medical imaging briefly. The current review is much more comprehensive and our approach is different, wherein we divide medical imaging applications into four problems that medical professionals can easily relate to.

We divide deep learning models into two groups as shown in Figure 1.1. Each of these is further subdivided into two and discussed in the next sections. We took this approach in order to organize the different deep learning techniques by their characteristics and distinguish between potential applications. The rest of this report is as follows. First we discuss the deep learning techniques briefly and move on to major applications of deep learning in the medical image analysis pipeline. We present applications in segmentation, registration, feature extraction and classification using deep learning. We then proceed to discuss the use of deep learning in fusing information from multiple models of data, which is typical of medical imaging. Finally, we discuss architectural refinements to adapt deep learning algorithms to specific medical imaging applications.

2 Deep Learning Models

Deep learning models can be divided into two groups according to their nature. Generative model can be considered as a probabilistic model of all variables whereas discriminative model provides a model relevant for the target variable conditional on the observed variables. Hence, a generative model can generate values of any variable in the model, while a discriminative model can only perform sampling of the target variables conditional on the observed quantities. In deep learning applications, these two classes of models can be considered complementary and can be used simultaneously in certain circumstances.

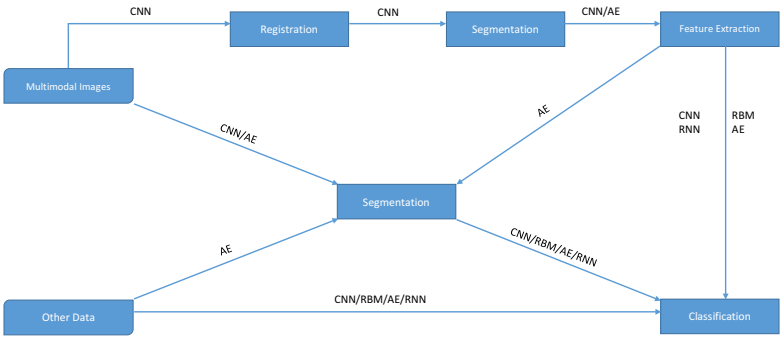

Figure 1.2: Applications of Deep Learning Techniques in Medical Imaging. The following abbreviations; CNN, RBM, AE, RNN; are used to denote convolutional neural networks, restricted Boltzman machines, autoencoders and recurrent neural networks respectively. We have included anomaly detection, progression tracking and recognition under the heading Classification in this illustration.

2.1 Discriminative Models

Discriminative classifiers model the posterior $p(y|x)$ directly or learn a direct map from inputs x to the class labels [10]. We first discuss convolutional neural networks (CNNs) and then recurrent neural networks (RNNs) in this section.

2.1.1 CNN

Convolutional Neural Networks (CNN) have been explored in the past and draw inspiration from typical neural networks [11]. The connectivity pattern of neurons in a CNN is inspired by the organization of the animal visual cortex. The architecture of CNN make the explicit assumption that the inputs are images that enables encoding certain properties into the architecture, making the forward function more efficient to implement and vastly reduce the amount of parameters in the network. While they were initially shown to be excellent at hand written digit recognition, the inability to scale CNNs to handle larger image sizes made them impossible to use in most applications. However, as this was largely due to hardware and memory constraints, coupled with a lack of sufficiently large datasets, recent advances in GPU computing and curation of large datasets such as ImageNet [12] have made it possible to use CNNs again. We aim to briefly look at the building blocks of CNNs and discuss two popular architectures while leaving detailed descriptions to other reports [13].

2.1.1.1 Convolution

The issue with traditional fully connected neural networks when dealing with images has always been the explosion of parameters when modeling each pixel as a single input node. For example, when we consider an image of size 100 x 100, we would have 10,000 input nodes which, in turn means 10,000 x 10,000 = 100 million parameters, if we have 10,000 nodes in the first hidden layer. As the networks become deeper, the number of parameters grows exponentially and nearly impossible to be handled even by the most advanced hardware. Instead of considering the whole image as input, an option is to learn a set of convolutional filters of varying sizes that considers a small image neighborhood at a time, which is much more tractable. The added advantage of this approach is that we can take spatial characteristics of the image into consideration as opposed to conventional neural networks. CNNs can be thought of as regular neural networks with two constraints [14]:

- Local Connectivity: In essence, each neuron is only connected to a small part of image as opposed to the whole image, as in regular neural networks.
- Parameter Sharing: Since the same convolution filter is applied across the image, weights between these filters maybe shared.

The input to the first convolutional layer may be an array of size l x w x n where l and w are the length and width of the image respectively while n denotes the number of channels. Natural images have three channels known as RGB which denote red, blue and green. Medical images on the other hand typically have a single channel known as gray-scale images. The output of this layer may be n_1 number of feature maps of size l_1 x w_1. The size of the output feature maps can be implicitly defined by the size of the convolutional filter

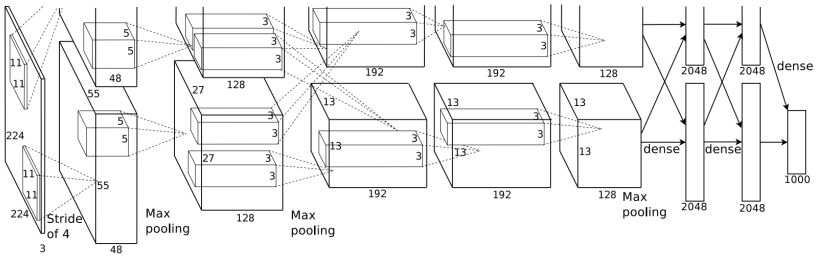

Figure 2.1: An illustration of the architecture of a CNN. This was first implemented using two GPUs and therefore, two parallel architectures are shown delineating the responsibilities between two GPUs. Illustration and description adapted from [16]

while the number of feature maps n_1 is explicitly defined as a hyper-parameter. These individual filters create a trainable mapping from the input feature map to the output feature map.

2.1.1.2 Pooling

A pooling operation essentially reduces the size of the activations for the next layer enabling us to use a smaller number of parameters progressively. CNNs use different types of pooling depending on the architecture but max-pooling is the most used pooling technique. If we consider an nxn region, max-pooling will replace that region with its max value reducing the size by a factor of n^2. Providing a small degree of spatial invariance can be considered as an added advantage of pooling.

2.1.1.3 Non-linearity

Since a cascade of linear systems (such as convolutions) generate another linear system, non-linearities between convolutions are added to expand the expressive power. Modern CNNs typically use ReLu non-linearity which can be expressed as $ReLu(x) = max(0, x)$. CNNs with ReLu non-linearity are shown to converge faster [15].

2.1.1.4 AlexNet

Adding the building blocks of CNNs together, we discuss AlexNet as the first example [16]. AlexNet was trained on ILSVRC 2012 training data which contained 1.2 million training images categorized into 1000 classes. AlexNet has 7 layers consisting of combinations of convolution, pooling and non-linearity as shown in Figure 2.1. Visualization of the output of layers as demonstrated by Figure 2.2 shows that earlier layers tend to learn low level features similar to Gabor-like oriented edges and blob-like features, while later layers tend to learn higher level features such as shapes and textures. Final layers appear to learn semantic attributes like eyes or wheels [17].

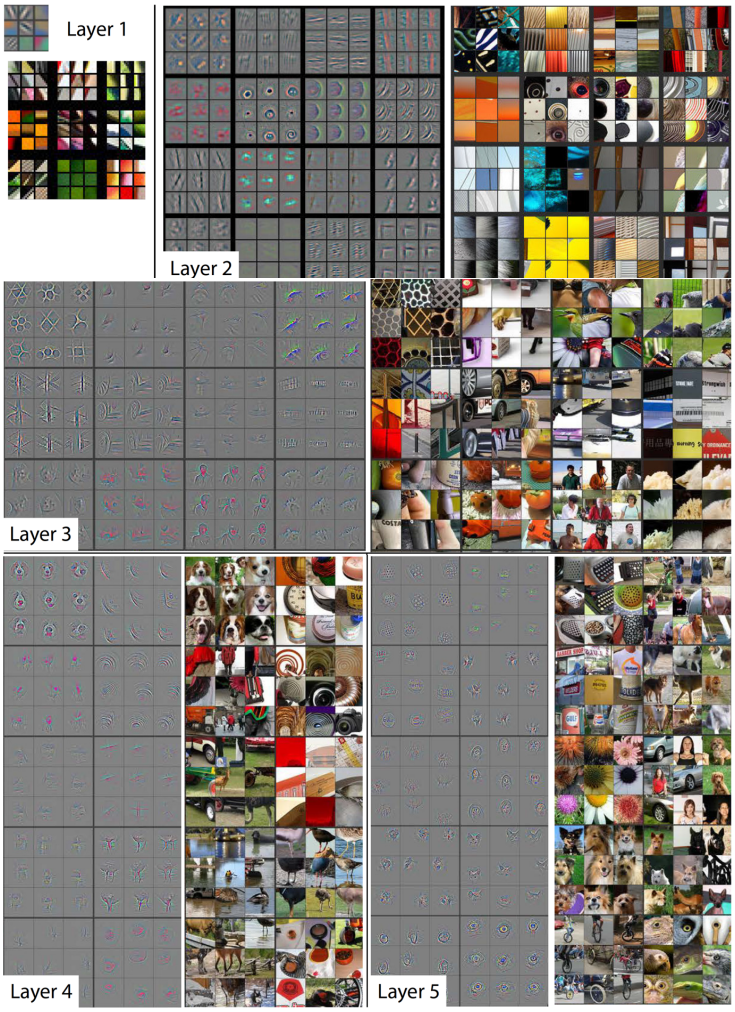

Figure 2.2: Visualization of features in a fully trained model. For layers 2-5, the top 9 activations in a random subset of feature maps across the validation data are shown, projected down to pixel space using deconvolutional network approach. The reconstructions are not samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map the corresponding image patches are also shown. Note: (i) the strong grouping within each feature map, (ii) greater invariance at higher layers and (iii) exaggeration of discriminative parts of the image, e.g. eyes and noses of dogs (layer 4, row 1, cols 1). Illustration and description adapted from [17].

2.1.1.5 GoogleNet

GoogleNet is another CNN that was trained on ILSVRC14 dataset and has 22 layers [18]. For details of the 22 layer architecture, readers are encouraged to refer elsewhere [18] which also contains a visualization of the network. They present a new architecture called inception that tries to use readily available dense components to approximate optimal local sparse structure of a convolutional vision network. A better description of the layers and rationale of their individual use can be found in the original paper. Applications of GoogleNet are discussed in the ensuing sections.

2.1.2 RNN

Despite the power of standard neural networks, one limitation they have is the assumption of independently generated samples. If the samples used to train a classifier are related in time or space, neural networks can fall short. A typical example is time-series data such as frames from a video or snippets of audio. Another issue with these types of data is that the samples may be vectors of different lengths at different time points, whereas typical neural networks rely upon the samples being vectors of fixed length. Recurrent neural network (RNN) was proposed to alleviate these issues, and it is a connectionist model with the ability to selectively pass information across a sequence, while processing sequential data one element at a time [19].

A standard RNN computes the hidden vector sequence $h = (h_1, h_2...h_T)$ and output vector sequence $y = (y_1, y_2, ...y_T)$ given an input sequence $x = (x_1, x_2, ...x_t)$ by iterating over the equations below from $t = 1$ to T [20]:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.1}$$

$$y_t = softmax(W_{hy}h_t + b_y) \tag{2.2}$$

Here W_{xh} is similar to a conventional weight matrix between the input and the hidden layer, while W_{hh} can be thought of as the weight matrix between the hidden layer and itself at adjacent time steps [19]. This can be represented in Figure 2.3 and the dynamics of the network across time steps can also be visualized by unfolding it as shown in Figure 2.4.

RNNs are better described elsewhere [19, 20, 21]. A popular extension of RNN is long-short term memory networks (LSTM) that are widely used today [21]. Currently, RNNs are rarely used in medical imaging although there may be interesting applications, as medical imaging applications do deal with sequential data.

2.2 Generative Models

Generative classifiers try to learn a model of the joint probability, $p(x, y)$, of the inputs x and the label y in order to make their predictions by calculating $p(y|x)$, and then picking the most likely label y [10]. Simply put, a generative model is used to specify a joint probability distribution over observations and labels. In the context of deep learning, two generative models are relevant, namely AutoEncoders(AE) and Restricted Boltzman Machines(RBM).

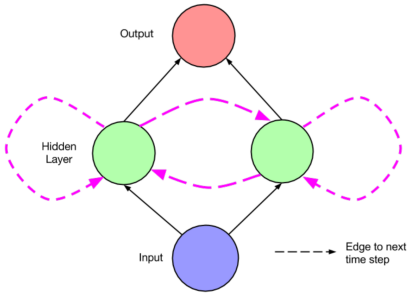

Figure 2.3: A simple recurrent network. At each time step t, activation is passed along solid edges as in a feed-forward network. Dashed edges connect a source node at each time t to a target node at each following time $t + 1$. Illustration adapted from [19]

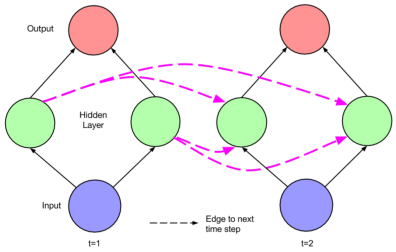

Figure 2.4: The recurrent network of Figure 2.3 unfolded across time steps. Illustration adapted from [19]

2.2.1 AE

Auto-encoder is a type of artificial neural network that can be defined with three layers: (i) input layer (ii) hidden layer and (iii) output layer. They transform inputs into outputs with the least possible amount of distortion. Auto-encoders were first introduced in the 1980s and their history and evolution are elaborated elsewhere [22]. The typical architecture of an AE is shown in Figure 2.5. It is predominantly an unsupervised learning algorithm but recent advances have made it possible to use a set of auto-encoders stacked on top of each other as a supervised learning algorithm [23].

Let us denote the input vector by $x \in \mathbb{R}^{D_I}$, where D_H and D_I denote the number of hidden and input units respectively. An auto-encoder creates a deterministic mapping from input to a latent representation y such that $y = f(W_1 x + b_1)$. This is parameterized by the weight matrix $W_1 \in \mathbb{R}^{D_H x D_I}$ and the bias vector $b_1 \in \mathbb{R}^{D_H}$. This latent representation $y \in \mathbb{R}^{D_H}$ is mapped back to a vector $z \in \mathbb{R}^{D_I}$ which can be considered as an approximate reconstruction of the input vector x with the deterministic mapping $z = W_2 y + b_2 \approx x$ where $W_2 \in \mathbb{R}^{D_H x D_I}$ and $b_2 \in \mathbb{R}^{D_I}$.

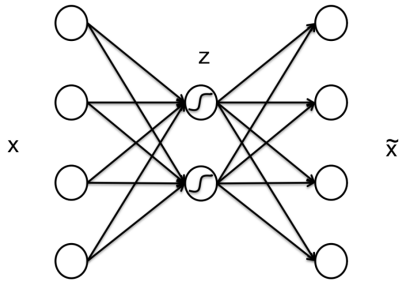

Figure 2.5: A typical AE transforms the input x to output $\widetilde{x}$ with minimum amount of distortion by encoding the input into z and decoding it back. Illustration adapted from [24].

2.2.2 RBM

While restricted Bolztman machines have been used in numerous applications, their most important use is as building blocks of deep belief networks [25]. They are used to learn important aspects of an unknown probability distribution based on samples from that distribution [26]. RBM consists of m visible units $V = (V_1, V_2..., V_m)$ that represent the observable data and n hidden units $H = (H_1, H_2, ...H_n)$ capturing the dependencies between observed variables [26]. The random variables (V, H) takes values $(v, h) \in 0, 1^{m+n}$ in a binary RBM. The joint probability distribution of such a RBM is given by the Gibbs distribution $p(v, h) = \frac{e^{-E(v,h)}}{Z}$ where

$$E(v, h) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

w_{ij} is a real valued weight associated between V_j and H_i for all $i \in 1, ..., n$ and $j \in 1, ..., m$. An illustration of a simple RBM is shown in Figure 2.6. It should be noted that an RBM only has connections between hidden and visible units but not between two variables of the same layer.

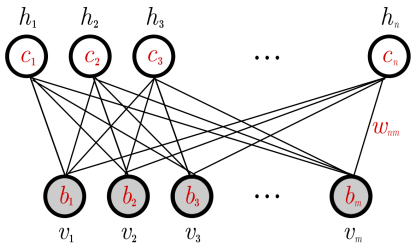

Figure 2.6: The undirected graph of an RBM with n hidden and m visible variables. Illustration adapted from [26]

3 Segmentation and Registration

3.1 Segmentation

Image segmentation is a process that divides image pixels or voxels into several subsets. Thus, the inputs are different types of images, whereas the outputs are groups of pixels with different labels. The goal of segmentation is to decompose an image into several parts or focus only on regions of interest(ROI), which helps with better understanding of the original image. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. In medical image segmentation, the major goal is to study anatomical structure, identify ROI, measure tissue volume to measure growth of tumour and help in treatment planning. The following subsections are organized in the following order. First, we talk about the state of art in biomedical image segmentation. Then we list recent studies based on deep learning algorithms. Finally, we discuss future work that could be done in this field.

3.1.1 Current status

As it is a fundamental problem, abundant research has been performed on image segmentation. These approaches could be mainly divided into five categories as mentioned elsewhere [27, 28]: Threshold based methods, Region based methods, Learning (classification) based methods, Model based methods and Atlas based methods.

i Threshold based methods use a threshold to determine the label of a pixel and divide the image using different labels. Here, the characteristics of pixels could be valued both in the spatial and wavelet domains. Threshold based methods are the simplest and fastest methods, but are sensitive to noise and threshold values [27, 28].

ii Learning(classification) based methods assume that the segmentation system can be trained by annotated data. Each pixel/voxel in the original image is identified as a certain label in the training data. We train the system with these annotations. New inputs are then predicted by the well-trained system [27, 28].

iii Region based methods are based on the principle of homogeneity - pixels with similar properties are clustered together to form a homogeneous region. There are 3 types of region based methods: region merging, region splitting and 'split and merge'. Region based approaches are powerful but they still suffer from under and over segmentation. Currently, some algorithms that combine region based and edge detection based methods have been developed to handle over/under segmentation problems [27, 28].

iv Model based methods assume that the structure of organs has a repetitive form of geometry and can be modeled probabilistically for variations of shape and geometry. Model based methods of segmentation involve active shape and appearance models, deformable models and level-set based models. These methods are relatively accurate. However, the computational complexity is high and manual interaction is required to imitate the model [29].

v Atlas based methods compile information on anatomy, shape, size, and features of different organs and soft tissues in the form of an atlas or look up table (LUT). Atlas guided approaches are similar to correlation approaches and the advantage of atlas-based approaches is that they perform segmentation and classification in one go. These methods are efficient and accurate. However, they face limitations in segmenting complex structures with variable shape, size and other properties and expert knowledge is required in building the database [30].

3.1.2 Segmentation by Deep Learning

In the biomedical field, deep learning algorithms often suffer from limited availability of data. In the early stages, only applications with big data sets can be found. Carneiro al. et [31] use Deep Belief Network (DBN) to segment the Left ventricle of the heart from ultrasound images, where the data contains 400 annotated images.

Later, several attempts were made to deal with small data size. One approach is to fix the low level features to reduce the burden of parameters training. Malon et al. [32] used man-made low-level features in their work. They then modified the weights and thresholds of low level features with convolutional neural networks (CNN). Bar al. etc [33] use the low level features of ImageNet, a well-trained network. This idea is inspired by the fact that different objects can be represented by the same low level features [34]. Another attempt is to decompose the original data into small patches. If the number of target segments in one image is large, we can divide the original image into small patches that contain only one segment each. A typical example is cell segmentation, such as electron microscopic images in the 2012 ISBI challenge. Ciresan et al. [35] used the classical CNN as a pixel classifier to segment electron microscopy images.

A third type of solution is the so-called U-net [36] architecture which is inspired by the fully convolutional network [37]. The main change here is the use of the up-sampling (up convolutional) layer, which helps to localize a segment and improve the resolution. There is no fully connected layer in the U-net architecture, since the network is trained by zooming in and out with convolutional layers. Milletari et al. [38] applied a modified version of U-net, namely V-net, to prostate segmentation. Instead of making use of 2D slices, they input volumetric data directly to their network. Brosch et al. were inspired by the idea of up-sampling(deconvolution) in their work [39, 40]. They developed an encoder network to segment Multiple Sclerosis lesions. This indicates that more biomedical segmentation applications could be tackled with the development of a common segmentation tool based on deep learning.

Another branch of biomedical segmentation utilizes different types of CNN. Many medical image segmentation tasks [41, 42] and [43] have been attempted on classical deep learning framework. Also Zhang et al. [44] utilize CNN to segment infant brain tissue into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). With the development of the deep learning algorithm, other extensions of CNN, for instance GooleNet, Alexnet, Segnet and ConvNet, have been introduced into biomedical territory. Maxout layer has been added between convolutional layer and max pooling layer to handle non-linear problems [45]. Roth et al. [46] make use of ConvNet to segment pancreas CT images. Compared to classical CNN, ConvNet has two more fully connected layers and

two drop out layers. Gaonkar et al. [47] suggest an object-to-object segmentation structure, which represents objects instead of features in each layer. This modification enables the inclusion of anatomical relationships into the segmentation.

There are also several works based on other methods [48, 49]. For instance, Liao et al. [48] applied Independent Subspace Analysis (ISA) Network on prostate segmentation. Vaidhya al. etc [49] tried to reconstruct the input by applying a stacked auto-encoder before segmenting brain tumours.

3.1.3 Future Work

Most biomedical image segmentation methods based on deep learning are mainly based on pixel classification. These methods are highly sensitive to the training data, and the results could be uncertain when we apply the well-trained system to a new dataset or data from a new machine.

It is a natural progression from pixel based classification to other methods, such as model-based methods and atlas-based methods. Deep learning algorithms can be utilized to fix or refine the parameters in model based methods. Also, we may define the similarity between input and atlas images with Deep Belief Networks or Auto-encoders. What is more, other recent work on deep learning can be brought into the medical image segmentation field, such as region-based convolutional networks [50] and long short term memory based image segmentation [51].

3.2 Registration

Image registration is about determining a spatial transformation (or mapping) that relates points in one image, to corresponding points in one or more other images [52]. The aim of biomedical image registration is to improve the results by combining images from different modalities or those captured at different times. A registration algorithm can be decomposed into two components: the similarity and the transformation model.

Similarity works as an evaluation of how well two images match each other. There are mainly two approaches to measure the similarity: geometric and intensity based approaches. Geometric approaches make use of anatomical features, such as landmarks, curves and surfaces, which help to identify the differences. Intensity based approaches match intensity patterns in each image using mathematical or statistical criteria. Intensity based approaches model the problem as a scoring problem. First, they define some similarity measurement, such as squared differences in intensities, correlation coefficient or measures based on optical flow and mutual information. Then they try to find the highest score on that measurement between two images, which indicates the best registration deformation.

The transformation model defines the way in which an input image should be deformed to match a source image. Mainly there are six types of transformation models: rigid model, spline model, elastic model, viscous fluid model, 'demons' model and finite element model. Rigid models handle the problem of translations, rotations and scaling between a moving image and target image. The most widely used spline models are B-spline [53] models and thin-plate spline models [54]. These models fix some landmark points and deform near the

fixed points. Elastic models[55] treat the source image as a linear, elastic solid and deform it using forces derived from an image similarity measure. In a fluid model [56], the deforming template image is considered as a viscous fluid whose motion is governed by its Navier-Stokes equation of conservation of momentum. 'Demons' [57] methods try to merge the moving image into the static image by the action of effectors, called demons, situated in these interfaces. Finite element models [58] divide an image into cells and assign to these cells a local physical description of the anatomical structure.

Until now, there is very limited work performed on medical image registration using deep learning techniques. The first attempt was by Wu [59] in 2013. In order to avoid the limitations of hand-crafted features, they attempt to take advantage of the unsupervised learning ability of deep learning networks. In their work, Independent Subspace Analysis (ISA) networks are used as a feature extraction method, which enables them to extract intrinsic features from the data automatically. Later, they introduced convolutional stacked autoencoder (SAE) [60] to extract 3D features for registration directly from image data. Here, the encoder is used to represent 3D features as a combination of 2D features. Then they use a decoder to reconstruct 3D patches from 2D features. Both encoder and decoder are SAEs in their work. Zhao [61] used CNN to find the rotation of rigid image registration. They attempt to divide image registration into global (rigid) and local(non-rigid) deformations. For rotation information in global deformation, they apply CNN to find 360 different angles of rotation. A brand new similarity measure based on deep learning algorithm has been described [62]. Cheng et al. [62] propose to judge the correspondence between two image patches with a binary classifier trained by a 3 layer deep neural network. This application inspires us to think about the possibility of other methods based on deep learning to measure similarity. There is still no deep learning technique based work on transformation models. This could be an interesting future direction.

4 Feature Extraction and Classification

Deep learning at its core is mainly a machine learning algorithm that has been sufficiently modified and optimized for performance. Any classification problem in machine learning can be modeled in two stages; feature extraction/selection and classification. The performance of the system is as good as the features that are used to train the system. It has long been established that handcrafting features and then performing feature selection tends to improve the performance of a learning system [63]. The allure of deep learning is the promise of letting the algorithm learn the feature representation from raw data instead of handcrafting features. Hence we shall review the deep learning techniques used for feature extraction and classification in medical imaging applications. It should be noted that these two sections have inherent overlaps.

4.1 Feature Extraction

As this report focuses on medical imaging applications, we restrict our review of feature extraction to that of images. Feature extraction may be regarded as one of the central problems in computer vision. Essentially it deals with extracting

'meaningful' descriptions from the images or image sequences. These descriptions can then be used for further processing such as registration, segmentation and classification. It used to be the case that 'meaningful' was dependent on the domain in question and hence various handcrafted features needed to be engineered for different applications. Deep learning however attempts to overcome this problem with the promise of letting the algorithm derive 'meaningful' descriptions from the raw data; images in this case. A good review of handcrafted features and their use cases can be found [64]. Medical imaging goes a step ahead of classical computer vision by further processing the feature descriptors to fit different modalities that can be found in medical images. For example, conventional computer vision descriptors like edge detectors can be used to identify and isolate edges in an image. These edge detectors can be used to identify the boundary between grey matter, white matter and cerebrospinal fluid in brain images, which is subsequently used to come up with a feature called 'grey matter volume'. In contrast, deep learning can be used to come up with feature descriptors by itself without the operator needing to handcraft the features. This has been demonstrated in [17] where initial convolution layers tend to learn low level features such as edge detectors, while later convolution layers tend to learn higher level features such as shape based features or texture based features that are directly relevant to the application in question. When looking at deep learning based feature selection, we can see two classes of techniques. The first technique uses generative models for feature representation while the second technique uses discriminative models. A concise description of these techniques is included and their applications in medical imaging are explored.

4.1.1 Generative Model based Feature Extraction

Autoencoders (AEs) are mostly used in an unsupervised manner in order to learn a representation for a set of raw data and has been used for the purpose of dimensionality reduction in the past [65]. It was first introduced in the mid 1980s by Rumelhart et al. [65] to address unsupervised backpropagation using input data. In the mid 2000s, AEs were proposed as dimensionality reduction techniques and subsequently have been heavily used as deep AE networks under the deep learning banner. The typical architecture stacks individually trained layers of AEs on top of each other and then fine-tunes the integrated network in a supervised manner. This is known as unsupervised pre-training and supervised fine-tuning in deep learning nomenclature.

An AE is essentially a feed forward neural network with an input layer, an output layer and one or more hidden layers in between them. An AE is trained to reconstruct its own inputs at the output layer and therefore, considered as an unsupervised learning model. Two typical parts in an AE are the encoder and decoder. They may be defined as transitions ϕ and ψ such that:

$$\phi : \mathcal{X} \rightarrow \mathcal{F} \tag{4.1}$$

$$\psi : \mathcal{F} \rightarrow \mathcal{X} \tag{4.2}$$

$$\arg\min_{\phi,\psi} \|X - (\psi \circ \phi)X\|^2 \tag{4.3}$$

Let us take the simplest case where there is one hidden layer. An AE takes the input $\mathbf{x} \in \mathbb{R}^d = \mathcal{X}$ and maps it onto $\mathbf{z} \in \mathbb{R}^p = \mathcal{F}$:

$$\mathbf{z} = \sigma_1(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{4.4}$$

This representation is known as code or latent representation. σ_1 is an element-wise activation function such as sigmoid function or ReLu. The next step is the mapping of z onto the reconstruction $\mathbf{x}'$ that has the same shape as x:

$$\mathbf{x}' = \sigma_2(\mathbf{W}'\mathbf{z} + \mathbf{b}') \tag{4.5}$$

AEs are also trained to minimize reconstruction errors:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2 = \|\mathbf{x} - \sigma_2(\mathbf{W}'(\sigma_1(\mathbf{W}\mathbf{x} + \mathbf{b})) + \mathbf{b}')\|^2 \tag{4.6}$$

Depending on the characteristics of the feature space $\mathcal{F}$, there are two types of AEs. If the feature space has lower dimensionality than the input space, then the AE has learnt a compressed representation of the input. Conversely, if the feature space has a higher dimensionality than the input space, then the AE has learnt a sparse representation (also known as overcomplete representation) of the input. This creates the potential for the hidden layers to learn the identity function which is not useful. However, several techniques have been proposed to overcome this restriction and their ability to learn a useful feature representation has been demonstrated.

A good example of the use of sparse autoencoders for feature learning can be found [66]. Kellenberg et al. use sparse autoencoders to learn features in an unsupervised way at multiple scales for mammography risk scoring. The use of sparse AEs has made it easier for them to interpret the learned features and also made the process cost-efficient and robust to noise. Their proposed sparsity regularizer is an amalgamation between the popular population sparsity and lifetime sparsity [66]. Liu et al. [67] also use sparse AEs to derive a better feature representation. Their approach can be used to demonstrate an alternative technique that actually performs an initial feature extraction step. They derive grey matter volumes from brain MR images and feed this as an input to the sparse autoencoders. They train the network one layer at a time as Bengio et al. suggested (this is also known as greedy layer-wise learning) [68] and use a softmax layer as the output layer in order to perform classification. The advantage of adding a softmax layer is the ability to do supervised fine-tuning [69]. A comprehensive review of the unsupervised pre-training and supervised fine-tuning approach can be found [70, 71]. A similar approach can be seen [72] clearly demonstrating the advantages of pre-training. Suk et al. also represent another application of autoencoders where AEs are used for fusing features from different modalities. We will discuss feature fusion in detail in section 5. Ithapu et al. [73] propose an imaging based enrichment criterion for clinical trials for mild cognitive impairment where AEs are used to come up with discriminative biomarkers.

The second type of generative model is RBMs. A deep network consisting of layers of RBMs is also known as a Deep Belief Network (DBN). An RBM can learn a probability distribution over its set of inputs. It is a type of Markov random field that can model data distribution, parameterizing it with the Gibbs

distribution over a bipartite graph between visible v and hidden variables h [74]: $p(v) = \sum_h p(v,h) = \sum_h \frac{1}{Z} e^{-E(v,h)}$, where $Z = \sum_v \sum_h e^{-E(v,h)}$ is the normalization term (the partition function) and $E(v,h)$ is the energy of the system. Assuming the RBM has m visible variables $V = (V_1, ...V_m)$ and n hidden variables $H = (H_1, ..., H_n)$, the energy function E can be written as:

$$Z = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i \tag{4.7}$$

where w_{ij} is a real valued weight associated with the edge between units V_j and H_i and b_j and c_i are real valued bias terms associated with the jth visible and ith hidden variable respectively. A comprehensive overview can be found [74].

Plis et al. demonstrate the use of RBMs in neuroimaging in a validation study. They use DBNs created from RBMs to come up with a better feature representation from brain fMRI and structural MRI data. This work is particularly enlightening for their modeling and use of raw MR data instead of processed data in order to come up with a feature representation. More applications of DBNs can be seen in Section 5 as they are also used as feature fusion techniques.

4.1.2 Discriminative Model Based Feature Extraction

In this report, we only consider convolutional neural networks (CNN) as a discriminative feature extractor. This approach is getting popular as it is relatively straightforward and lifts the burden of training a CNN from scratch for the most part. That also means expert knowledge on optimizing CNNs does not become a restriction and therefore the applicability of this approach is wide. The typical analysis pipeline starts with a pretrained CNN that is used to extract 'meaningful' features from raw data. This CNN may have been trained on a completely different dataset such as ImageNet [12] although if the CNN was trained on a similar dataset, the transferability of knowledge improves [75]. These extracted features are then used for classification using conventional classification techniques and may or may not be used alongside handcrafted features.

A good example of this approach can be drawn from the works of Bar et al. [76] where they use a CNN pretrained on ImageNet dataset for chest pathology detection using chest X-rays. Their approach is straight forward and uses the CNNs as a feature extractor where activations from convolution layers are considered as features and a SVM [77] is used to train on these features and some handcrafted features. Another approach proposed by Ginneken et al. [78] use OverFeat CNN trained on object detection on natural images to detect pulmonary nodules using CT images. Their classification was also performed using SVMs and CNN was used as a feature extractor. Arevalo et al. [79] also use AlexNet trained on ImageNet as a feature extractor, and train an SVM along with a range of handcrafted features for mammography mass lesion classification.

4.2 Classification

Classification is the problem of distinguishing the category (or class) of a new observation with a model that was trained on a set of data containing obser-

vations whose categories are known. The simplest classification problem is a binary classification problem where only two categories are available. A typical example would be diagnosing a patient with prostate cancer using a prostate MR image. In this case, the model would have been trained on a training dataset with patients who have prostate cancer (positive examples) and patients who do not have prostate cancer (negative examples). When a patient undergoes a MRI scan, it is considered as a new observation, and the model is consulted to identify whether the patient has cancer or not. Classification is an instance of supervised learning where a training set with known categories is available to train the model. A good overview of conventional supervised classification techniques can be found [80]. The practical utilization of CNNs in classification is two fold. The first is to design and train the CNN from scratch using data. The second is using a network that was already trained and further optimizing it to the application. This is also known as transfer learning. We discuss both approaches and their relative advantages and disadvantages in an effort to demonstrate their use in medical imaging. Tajbakhsh et al. [81] recently discussed these approaches and their practical relevance.

4.2.1 Designing and Training from Scratch

In one of the earliest publications using CNNs to detect lung nodules, Lo et al. [82] propose the use of a basic CNN that is not as deep as the ones that are used today but given that it was in 1995, this was to be expected, as CNNs only became deeper as the computational power increased. They also introduced data augmentation approaches by making use of rotation and shift invariance characteristics of medical images in order to enlarge their training dataset. The CNN proposed by Sahiner et al. [83] in 1996 enabled the use of full images as input instead of the region of interest (ROI) used by Lo et al. Their network was more generalized compared to Lo et al., but still shallow by current standards for the same computational reasons.

With the advent of GPU computing, medical imaging has seen a new wave of CNN applications recently. Ciresan et al. [84] use CNNs to detect mitosis using the publicly available MITOS database [85]. A typical problem faced by researchers who train their own network from scratch is the lack of accurately annotated data that is available in other domains such as ImageNet [12]. Ciresan et al. were able to overcome that problem as they were using pixels as singular units, therein generating millions of effective training instances although there were only 50 images available. They also exploit rotational invariance to enlarge their training set. Tajbakhsh et al. [86] use an ensemble of CNNs to detect polyps from colonoscopy videos. They use a network that they had trained from scratch previously as a feature extractor, and create an ensemble of CNNs for polyp detection. The use of ensembles can mitigate overfitting which is typical of medical imaging as the available datasets are small. Sarraf et al. [87] propose a dual CNN based classifier system that can be used to identify patients with Alzheimer's Disease using structural and fMRI data. Their preprocessing techniques are of particular interest, as they directly lead to the network's ability to fuse data from two modalities. Roth et al. [88] propose a new data augmentation technique in order to enhance the training data when training a CNN from scratch. They identify it as 2.5D decomposition in representing 3D images.

So far we have looked at 2D convolution where the convolution operator iterates along the width and height of an image. However, medical imaging modalities typically generate volumetric images which are actually 3D (ie: also have a depth apart from the width and height). 3D convolution was used by Anirudh et al. [89] where they train a 3D CNN for lung nodule detection. Their approach is particularly interesting with the region growing labeling system they propose where the radiologist only has to define the centre pixel of a nodule, and the rest of the nodule is labeled using a region growing technique in 3D space. Other work [90, 91, 92] exploit the inherent volumetric nature of medical imaging data in order to come up with 3D CNNs which are looked at in detail in Section 6.

4.2.2 Using Pretrained CNN

While training a CNN from scratch has its advantages, it is also not without complications. As mentioned earlier, medical imaging datasets typically do not have a large number of annotated training data, since expert annotation is expensive and the rate of data acquisition and dissemination is slow. Training a deep learning pipeline from scratch is also expensive on computational resources and extremely time consuming. In addition, deep learning becomes complicated with over-fitting and convergence issues, whereby hyperparameters constantly need to be optimized, which is a repetitive process.

The use of pretrained networks has been proposed as an alternative to training the CNN from scratch as it can alleviate the complications to a certain extent. This is also known as transfer learning. This approach can be divided into two major branches; the use of pretrained networks without further optimization for feature extraction or classification, and the use of pretrained networks as an initialization method and subsequently tailoring the network (fine-tuning) to the specific application in question [81]. The recent work by Azizpour et al. [75] asserts that the success of knowledge transfer in CNNs depends on dissimilarity between the database on which a CNN is trained and the database to which the knowledge is transferred. Although the dissimilarity between networks trained on natural image datasets like ImageNet [12] and medical imaging datasets is abundantly apparent, recent research demonstrates the success of this approach in certain circumstances. As already discussed, using a CNN as a feature extractor, we concentrate on fine-tuning pretrained networks to specific applications and draw upon several examples of such optimizations.

A prime example can be drawn upon from the work of Chen et al. [93] where the authors use a pretrained CNN on ImageNet and fine-tune the fully connected layers using their own data in an effort to optimize the CNN and report state-of-the-art performance for localizing standard planes in ultrasound images. Carneiro et al. [94] append a multinomial logistic regression layer to a pretrained CNN on ImageNet and use that to fine-tune the CNN for mammogram analysis. They also combine unregistered craniocaudal and mediolateral oblique mammogram views using a final CNN to create their classification system. Shin et al. [95] demonstrate an application where they use fine-tuned pretrained CNN to map semantic information to medical images using a set of radiology images and clinical reports. The flexibility of using a pretrained CNN is that one can decide which layers need fine-tuning and which layers can be used as is. In other work [96], the authors decided to fine-tune all layers of the pretrained CNN and

use it for automatic classification of interstitial lung diseases. Their approach of the attenuation rescaling scheme to produce 3-channel images out of 1-channel CT scans is of particular importance as CNNs pretrained on natural images typically require 3-channel images, as they were trained on 3-channel (RGB) images whereas medical images are typically 1-channel (grayscale).

5 Multimodal Fusion

As noted earlier and by Shin et al. [97], a fundamental challenge in deep learning in medical imaging is the paucity of data. While deep learning based models have often been shown to improve on the state-of-the-art where applied, there is no equivalent to ImageNet for medical imaging. Greenspan et al. [9] express the belief that this has prevented deep learning models in medical imaging from achieving the substantial 10% increase in performance that similar models have achieved elsewhere in computer vision.

At the level of developing a single model, Shin et al. [97] outline three strategies which researchers use to create deep learning models:

1. Use an off-the-shelf model and rely on transfer learning,
2. Train a model from scratch in spite of the limitations, or
3. Use features deep neural networks for feature extraction only.

These methods have been discussed in previous sections. We also do not cover image fusion as has been surveyed by James and Dasarathy [98].

In some cases, however there are additional sources of data available, and recent work has attempted to integrate multiple data sources to improve model performance. Often a patient has not one, but a series of medical imaging of different types. A patient may also have electronic medical records (EMR) or other test results available. Broadly, this approach can be termed multimodal fusion. In this section, we investigate the fusion of medical imagery with either other types of medical imagery or non-image data.

The use of different MRI modalities is particularly common, as it has demonstrated improvements in the diagnosis of disease such as prostate cancer [99]. During a single imaging session with MRI, it is possible to capture several different modalities such as T1, T2, diffused-weighted dynamic contrast enhancement, and spectroscopy. For more serious illnesses, MRI imaging may be done in conjunction with PET scanning as the former delivers better imagery on soft tissue while the later is clearer for hard tissue.

There is an existing body of conventional techniques that already make use of different image modalities, including a wide variety of models which compete in the MICCAI associated BRATS competition [100]. We only consider deep learning based models here.

5.1 Multimodal fusion in other areas of computer vision

The idea of training deep learning models with multimodal data is not new in machine learning. This work builds on a longer tradition of research in speech recognition which seeks to exploit both types of data such as [101]. Original work by Ngiam et al. [102] addressed the use of building multimodal models

from audio and video data in the context of deep learning in speech recognition. In their paper, the authors compare five architectures, three using RBMs and two using AEs. The three RBM models are:

1. Separately trained audio and video models
2. Shared representation RBM from concatenated audio and video input
3. Greedy layerwise trained models from (1) which are then combined with deep hidden layer and then trained again.

They then use the RBMs to train two different types of AE architectures for denoising, using either crossmodal training with video or bimodal training with video and audio. Their results show an accuracy improvement of 8% accuracy over contemporary state-of-the-art methods, when they use both modalities. These architectures are commonly adapted in multimodal medical imaging deep learning models.

Another landmark study in computer vision are by Srivastava and Salakhutdinov [103] who investigate the use of DBMs for the purpose of image retrieval using textual and image data and find that they outperform classical SVM and LDA approaches. An early use of temporal information is by Le et al. [104] who use multimodal image and time series data to study action recognition using independent subspace analysis (ISA) network to good success.

5.2 Multimodal fusion and deep learning in medical imaging

In the medical imaging domain, the first multimodal study using deep learning was done by Shin et al. [105] for the purpose of detecting the liver, heart, kidneys and spleen. Shin et al. [105] used DCE-MRI with $t = 40$ time points taken over 6 second intervals from 76 patients with liver and kidney metastases. Their model is a shallow stacked AE model defined with reference to [102], ie: their architecture does not include a shared layer. The features from the stacked autoencoders is combined with one-vs-all logistic regression on image patch samples to detect the appropriate organs. Their results compare favorably with features derived from HOG and DFT which are common image and time series features.

Work by Suk et al. [106] classify Alzheimer's disease (AD) and mild cognitive impairment (MCI) from non-affected patients. Their dataset comprises 297 patients with AD and MCI as well as 101 non-affected patients with two different imaging modalities: diffusion MRIs and PET scans. Their model is based around stacked Boltzmann machines in a so-called deep Boltzmann model (DBM) for learning hierarchical features which is then fed to a SVM for multiclass classification into diagnostic categories. Unlike Shin et al. [105], the architecture of Suk et al. [106] have a shared deep layer between the modalities. Although they note difficulty in training their model, the authors show improvement over previous work based around hand-crafted features on the same AD/MCI vs NA classification task.

Liu et al. [107]also use stacked AE for the prediction of AD and MCI. Their data comprises of 331 patients with imaging performed on T1 MRI and FDG-PETs, and their model follows the concatenated fusion model of Ngiam et

al. [102] which they train using a denoising technique where a proportion of data from one modality is corrupted. The features learned by the stacked AEs are then used in a soft-max layer to perform the classification. They find that their results offer a minor improvement over SVM-based methods. Li et al. [108] also examine the problem of classifying AD and MCI, but they use data fusion techniques prior to feature extraction with DBMs and do not compare their model to other methods in the literature.

In the area of lesion detection, Brosch, et al. [109] develop another stacked autoencoder model. Their dataset consists of 474 patients with secondary progressive multiple schlerosis from whom T1, T2, and Proton density weighted (PD) MRI was collected. Brosch, et al. primarily use their model for dimensionality reduction and do not compare their model directly with others, but simply measure the strength of the correlation with clinical diagnostic tests for each of the measures. They find that the results are highly significant, and hence their results are clinically significant. Their model was subsequently developed and is discussed elsewhere in this report.

As mentioned previously, the BRATS medical imaging dataset [100], has proven to be fertile ground for experimentation with CNNs by a variety of researchers. The BRATS-2013 dataset consists of 30 patients with high and low grade brain tumours and four MRI imaging modalities which are collected for each patient including: T1, T1ce, T2, and FLAIR. The image data has ground truth labels and voxels are segmented into 5 categories. The BRATS-2013 is not volumetric, and this allows Havei et al. [45] to use an architecture of stacked convolutional networks directly on the 4 registered modalities. Since then, a variety of research on tumour segmentation has integrated different MRI modalities using convolutions directly [110, 111, 112].

Recent work by Nie et al. [113] aims to predict patient survival time which is treated as a classification problem of either short or long. Their dataset consists of 61 patients with malignant and recurring or very aggressive (WHO III and IV) tumours who have the following modalities collected: T1 MRI, resting-state fMRI and DTI. Each modality is handled separately through 4 layers of convolutions and the resulting features are fused and passed to three fully connected layers. The architecture is inspired by [103].

An interesting example of image fusion is work by Xu et al. [114] which attempts to classify the severity of cervical dysplasia, a precancerous condition useful for the early detection of cervical cancer. Their dataset consists of a random sample of 690 visits from the Guanacaste project database, and includes Cervigram RGB imagery, Pap and HCV test results as well further non-image diagnostic data. Authors use a pretrained AlexNet CNN model whose output is concatenated with the preprocessed non-image data and passed through several fully connected ReLu layers before finally going to a soft-max layer. Moreover they also investigate the utility of the hidden joint layers, and find they improve the overall quality of the model. The resulting model significantly improves on the existing state of the art.

6 Architectural Refinements

This report discussed the adaptation of deep learning techniques stemming from computer vision and artificial intelligence communities in medical imaging so far.

Deep learning techniques in medical image analysis have achieved promising results on various applications, including the diagnosis of Alzheimer's disease and mild cognitive impairment [115], organ segmentations [105]and detection [116]. Although we briefly looked at the architectural refinements that have been proposed in different applications, we believe the different architectural refinements deserve a lengthier discussion. Medical imaging problems typically differ from conventional computer vision techniques and hence CV techniques need to be modified to suit specific medical imaging applications. In this section, we discuss the amendments to deep learning techniques proposed by various researchers to improve their applicability in medical imaging.

Suk et al. [115] use stacked auto-encoder (SAE)for classification of Alzheimer's, Mild Cognitive Impairment and Healthy Controls from target samples. The proposed system exploits the latent information existing among the features and concatenates it with the original features. The approach has yielded higher accuracy when compared to the prior methods [117] and [118]. Shin et al. [105]apply deep learning techniques for organ segmentation from MR images. The approach employs hierarchical feature clustering. The method also made use of probabilistic patch-based method for multiple organ detection. The proposed system uses SAE for organ detection. Results show that deep learning methods have significantly improved the accuracy in organ detection from MR medical images. The architecture of restricted Boltzmann machines (RBM) are exploited for lung texture classification and airway detection in CT images [116]. The unsupervised learning approach of RBM helps to learn from unlabelled data. The proposed approach uses a combination of generative and discriminative learning that outperforms traditional methods [119]. Automatic feature representation for medical image segmentation is studied using deep learning [120]. Multiple instance learning (MIL) is used for classification of abnormalities. The paper also compares the performances of weakly supervised and fully supervised learning approaches. Results show that weakly supervised learning is better than a fully supervised learning approach.

Liao et al. [121] apply stacked independent subspace analysis network (ISA) to learn features from prostrate MR images. The extracted features contain anatomical information which is used for automatic prostrate MR segmentation. Results show that the proposed technique outperforms prior segmentation methods [122]. Automated detection of bone lesions in CT image using convolutional neural network classifier is proposed [123]. During testing, per individual classification probability is computed. Experimental results show that the proposed system outperforms previous methods. Roth et al. [124] propose automated Lymph Node (LN) detection in CT images using deep convolutional neural networks. Prior methods on lymph node detection directly use 3D information. In the proposed method the LN CAD system has high sensitivities at the first stage and gradually reduces false positives. Experimental results show that the proposed system outperforms previous methods. Prasoon et al. [125] propose the novel approach of combining three 2D convolutional neural networks for knee cartilage segmentation in MR images. The proposed technique employs voxel classification and extracts only 2D features. Experimental results show that the method is better than the prior methods.

Jonathan Masci et al. [126] present an algorithm to speed up training of MaxPooling Neural Convolutional Networks (MPCNN) for image segmentation. The main idea of the algorithm is that the network is trained on whole images

rather than manifold image patches. This is implemented by adding the MaxPoolingFragment (MPF) layer with the back-propagation procedure computing the partial derivative of MPF layer output in accordance with its input. The proposed algorithm achieved outstanding results in decreasing training time. A customized CNN was designed for classification of lung images with interstitial disease by Li et al. [127]. The proposed CNN consists of a single convolutional layer, a max pooling layer, and three fully connected layers. In addition, a drop-out algorithm (random disabling neurons during training) enhances performance of the proposed system. The presented approach outperforms systems with prior feature extraction for the same task. Payan et al. [128] use sparse AEs and 3D CNN to classify brain MRI images. At first, features are extracted from 3D image patches using an autoencoder, which is a 3-layer NN. Subsequently, 3D CNN computes the conditional probabilities for each possible class (healthy brain, mild cognitive impairment, Alzheimers disease). The experiments show that 3D CNN slightly outperforms 2D CNN for the same data.

Roth et al. [129] apply CNN to computed tomography (CT) images. The aim is to classify images into 5 anatomical classes. The ConvNet architecture includes five convolutional layers and three fully connected layers. There are max pooling and drop-out operations between these eight layers. Data augmentation is performed to increase the validation dataset. Van Grinsven et al. [130] use CNN for hemorrhage detection in colour fundus images. The architecture of the network includes five convolutional layers, two max pooling layers and fully connected layer. A selective sampling strategy is proposed to speed up the training phase. Difficult training samples are presented to the network at first and larger weights are assigned to incorrectly classified samples with bigger errors(SeS algorithm). The training using the proposed strategy is approximately three times faster than the usual one. In addition, the overall performance of the CNN with SeS is higher than without the SeS selective sampling strategy. Bekker et al. [131] present multi-view NN (MV-NN) architecture for classification of breast micro calcifications using cranial-caudal (CC) and mediolateral-oblique (MLO) mammography views. The MV-NN classification model consists of two neural networks which are learned in parallel. Outputs of these networks are given to a single-neuron layer as input. The suggested method is evaluated on a large multi-view dataset. A deep learning framework for sub-cortical brain structure segmentation in MR images is proposed by Mahsa Shakeri et al. [132] A fully convolutional network takes 2D slices of 3D brain MR images as input, while previous methods were applied to image patches. Two different MRI datasets were used to validate the framework.

7 Open Problems

While we have discussed a number of problems that are effectively handled by deep learning techniques, we have not seen the same improvements in medical imaging that were apparent in computer vision with the advent of deep learning. We assert that this is because of two reasons: (i) lack of sufficiently large annotated datasets and (ii) issues with direct adaptability of deep learning techniques stemming from CV applications. The former may be handled by a number of ways such as increasing the number of data acquisitions and devoting more funding to acquire expert annotations, which are largely independent

from CV techniques. However, there are a few techniques that we can adapt to enlarge the datasets we already have and improve the annotations. Characteristics such as shift and translational invariance in medical images can be exploited to enlarge the datasets we already have. Autoencoders may also be used to generate similar pseudo data using the actual datasets we have. Another approach we can explore stems from the recent approach by Ledig et al. on generative adversarial networks (GAN) which can be used to improve the spatial resolution of the datasets we have which in turn enable us to use old low resolution datasets that are usually discarded. Ultimately, creating a unified repository that can be used to collect data from all over the world may help us in achieving the size of the datasets we are after. Annotating data has mostly been manual so far with the domain experts but recent work by Albarqouni et al. [133] have demonstrated the use of crowd sourcing to perform annotations which can go a long way in annotating the data we already have.

A number of researchers are exploring the issues with direct adaptability of deep learning techniques and are making significant progress as can be seen from this review. We believe transfer learning has made deep learning techniques more approachable to medical professionals as the learning curve and the effort needed to train a system is low. However, we also believe that transfer learning from a network that was actually trained on medical images may be more beneficial and to that end, we see room for improvement. We also believe typical deep learning libraries could be tailored to facilitate the inherent complexity of medical images which would make it more approachable to medical professionals. Another prospective improvement is a visualization library tailored for medical professionals as the method of volumetric visualization is different from typical two dimensional visualizations. While there are heuristics to determine the architecture of the deep learning technique of choice, it is still an active research area that we believe could benefit from a systematic method of architecture design including methods to tune relevant hyper-parameters.

8 Concluding Remarks

In this report we have briefly discussed state-of-the-art deep learning techniques and presented their applications in medical imaging. Historically, techniques from computer vision and machine learning have been adapted readily to medical imaging problems and we expect nothing short of deep learning techniques as well. However, deep learning techniques are continually evolving and have been considered as a black-box in most applications until recently, which has made it difficult to adapt to medical imaging tasks. Our effort here was to demystify the black-box and demonstrate potential applications in medical imaging. Our approach was to briefly describe the building blocks of deep learning techniques and direct the reader to better descriptions. We then reviewed the applications of deep learning in medical imaging depending on the analysis pipeline including registration, segmentation, feature extraction and classification. We have also have discussed multi-modal data fusion and the specific architectural refinements needed to adapt deep learning techniques to medical imaging applications. Finally, we present potential future work that can be carried out to make deep learning more attractive to medical professionals as well as the medical imaging community. By providing a critical review of the applications

of deep learning techniques in medical imaging, we believe we can aid to expand the boundaries and reach the level of success that deep learning has enjoyed in other fields.

Bibliography

[1] D. Ganguly, S. Chakraborty, M. Balitanas, and T.-h. Kim, *Medical Imaging: A Review.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 504–516. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-16444-6_63

[2] Y. Q. Zhang and J. C. Rajapakse, *Machine Learning in Bioinformatics.* Wiley, mar 2008, vol. 7, no. 1. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16761367

[3] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[4] M. Yang, Y. Yuan, X. Li, and P. Yan, "Medical Image Segmentation Using Descriptive Image Features," *Procedings of the British Machine Vision Conference 2011*, pp. 94.1–94.11, 2011. [Online]. Available: http://www.bmva.org/bmvc/2011/proceedings/paper94/index.html

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, may 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14539

[6] D. Y. Li Deng, "Deep learning: Methods and applications," Microsoft, Tech. Rep., May 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/

[7] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of Deep Learning in Biomedicine," pp. 1445–1454, may 2016. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.molpharmaceut.5b00982

[8] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine learning in genomic medicine: A review of computational problems and data sets," pp. 176–197, jan 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7347331/

[9] 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, may 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7463094/

[10] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 841–848.

[11] 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.

[12] S. V. Lab. (2016) Imagenet dataset. [Online]. Available: http://image-net.org/

[13] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu. (2016, jan) A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. [Online]. Available: http://arxiv.org/abs/1601.06615http://dx.doi.org/10.3389/frobt.2015.00036

[14] C. M. Bishop, "Pattern Recognition and Machine Learning," *Journal of Electronic Imaging*, vol. 16, no. 4, p. 049901, 2007. [Online]. Available: http://www.library.wisc.edu/selectedtocs/bg0137.pdf$\delimiter"026E30F$nhttp://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.2819119

[15] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proceedings of the 27th International Conference on Machine Learning*, no. 3, pp. 807–814, 2010.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[17] Y. Lecun and M. Zeiler. (2015) Visualizing and Understanding Convolutional Networks.

[18] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, "Going deeper with convolutions," *arXiv preprint arXiv: 1409.4842*, 2014. [Online]. Available: /citations?view{_}op=view{_}citation{&}continue=/scholar?hl=ja{&}as{_}sdt=0,5{&}scilib=1{&}citilm=1{&}citation{_}for{_}view=KtmM-dAAAAAJ:JV2RwH3{_}ST0C{&}hl=ja{&}oi=p

[19] Z. C. Lipton, "A Critical Review of Recurrent Neural Networks for Sequence Learning," *CoRR*, vol. abs/1506.0, pp. 1–38, 2015. [Online]. Available: http://arxiv.org/abs/1506.00019

[20] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech Recognition With Deep Recurrent Neural Networks," *Icassp*, no. 3, pp. 6645–6649, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp={&}arnumber=6638947{&}url=http{%}3A{%}2F{%}2Fieeexplore.ieee.org{%}2Fstamp{%}2Fstamp.jsp{%}3Ftp{%}3D{%}26arnumber{%}3D6638947

[21] M. Hermans and B. Schrauwen, "Training and Analyzing Deep Recurrent Neural Networks," *Nips*, pp. 190–198, 2013.

[22] P. Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," *ICML Unsupervised and Transfer Learning*, pp. 37–50, 2012.

[23] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16764513$\delimiter"026E30F$nhttp://www.mitpressjournals.org/doi/abs/10.1162/neco.2006.18.7.1527

[24] Q. V. Le, "A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks," *Tutorial*, pp. 1–20, 2015.

[25] G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines," *Computer*, vol. 9, no. 3, p. 1, 2010. [Online]. Available: http://learning.cs.toronto.eduhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.9573{&}rep=rep1{&}type=pdf

[26] A. Fischer and C. Igel, "An Introduction to Restricted Boltzmann Machines," *Lecture Notes in Computer Science: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, vol. 7441, pp. 14–36, 2012. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-33275-3{_}2$\delimiter"026E30F$nhttp://link.springer.com/chapter/10.1007/978-3-642-33275-3{_}2http://link.springer.com/chapter/10.1007/978-3-642-33275-3{_}2

[27] G. G. P. Kavitha. Survey on medical image segmentation and its methods.

[28] S. Masood, M. Sharif, A. Masood, M. Yasmin, and M. Raza, "A survey on medical image segmentation," *Current Medical Imaging Reviews*, vol. 11, no. 1, pp. 3–14, 2015.

[29] T. McInerney and D. Terzopoulos, "Deformable models," in *Handbook of Medical Imaging*. Academic Press, Inc., 2000, pp. 127–145.

[30] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical image analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[31] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 968–982, 2012.

[32] C. D. Malon, E. Cosatto *et al.*, "Classification of mitotic figures with convolutional neural networks and seeded blob features," *Journal of pathology informatics*, vol. 4, no. 1, p. 9, 2013.

[33] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 94 140V–94 140V.

[34] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 609–616.

[35] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

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

[37] 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*, 2015, pp. 3431–3440.

[38] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *arXiv preprint arXiv:1606.04797*, 2016.

[39] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, and R. Tam, "Deep convolutional encoder networks for multiple sclerosis lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 3–11.

[40] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[41] L. C. Garcıa-Peraza-Herrera, W. Li, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin. (2016) Real-time segmentation of non-rigid surgical tools based on deep learning and tracking.

[42] M. Havaei, F. Dutil, C. Pal, H. Larochelle, and P.-M. Jodoin, "A convolutional neural network approach to brain tumor segmentation," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, 2015, pp. 195–208.

[43] 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.

[44] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, mar 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25562829http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4323729http://linkinghub.elsevier.com/retrieve/pii/S1053811914010660

[45] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, 2016.

[46] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers, "Deep convolutional networks for pancreas segmentation in ct imaging," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 94 131G–94 131G.

[47] B. Gaonkar, D. Hovda, N. Martin, and L. Macyszyn, "Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2016, pp. 97 852I–97 852I.

[48] S. Liao, Y. Gao, A. Oto, and D. Shen, "Representation learning: a unified deep learning framework for automatic prostate mr segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 254–261.

[49] K. Vaidhya, S. Thirunavukkarasu, V. Alex, and G. Krishnamurthi, "Multi-modal brain tumor segmentation using stacked denoising autoencoders," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, 2015, pp. 181–194.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[51] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[52] W. R. Crum, T. Hartkens, and D. Hill, "Non-rigid image registration: theory and practice," *The British Journal of Radiology*, 2014.

[53] 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.

[54] F. Bookstein, "Thin-plate splines and the decomposition of deformation," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 10, 1988.

[55] I. N. Figueiredo, C. Leal, L. Pinto, P. N. Figueiredo, and R. Tsai, "An elastic image registration approach for wireless capsule endoscope localization," *arXiv preprint arXiv:1504.06206*, 2015.

[56] H.-H. Chang and C.-Y. Tsai, "Adaptive registration of magnetic resonance images based on a viscous fluid model," *Computer methods and programs in biomedicine*, vol. 117, no. 2, pp. 80–91, 2014.

[57] S. R. Chowdhury, R. Ray, N. Dey, S. Chakraborty, W. B. A. Karaa, and S. Nath, "Effect of demons registration on biomedical content watermarking," in *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on.* IEEE, 2014, pp. 509–514.

[58] H. Zhong, J. Kim, H. Li, T. Nurushev, B. Movsas, and I. J. Chetty, "A finite element method to correct deformable image registration errors in low-contrast regions," *Physics in medicine and biology*, vol. 57, no. 11, p. 3499, 2012.

[59] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of mr brain images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 649–656.

[60] G. Wu, M.-J. Kim, Q. Wang, B. Munsell, and D. Shen, "Scalable high performance image registration framework by unsupervised deep feature representations learning," *Transactions on Biomedical Engineering*, 2015.

[61] L. Zhao and K. Jia, "Deep adaptive log-demons: Diffeomorphic image registration with very large deformations," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[62] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–5, 2016.

[63] V. Kumar and S. Minz, "Feature Selection: A literature Review," *Smart Computing Review*, vol. 4, no. 3, 2014.

[64] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs hand-crafted features for pedestrian gender recognition," in *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, 2015, pp. 1263–1266.

[65] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[66] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm, "Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1322–1331, May 2016, bibtex: kallenberg_unsupervised_2016.

[67] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2014, pp. 1015–1018.

[68] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, and others, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[69] P. Lamblin and Y. Bengio, "Important gains from supervised fine-tuning of deep architectures on large labeled sets," in *NIPS*2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2007.

[70] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010. [Online]. Available: http://www.jmlr.org/papers/v11/erhan10a.html

[71] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-35289-8_26

[72] H.-I. Suk and D. Shen, "Deep learning-based feature representation for AD/MCI classification," *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 16, no. Pt 2, pp. 583–590, 2013.

[73] V. K. Ithapu, V. Singh, O. C. Okonkwo, R. J. Chappell, N. M. Dowling, S. C. Johnson, and Alzheimer's Disease Neuroimaging Initiative, "Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment," *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 11, no. 12, pp. 1489–1499, Dec. 2015.

[74] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2012, pp. 14–36. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-33275-3_2

[75] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of Transferability for a Generic ConvNet Representation," *arXiv:1406.5774 [cs]*, Jun. 2014, arXiv: 1406.5774. [Online]. Available: http://arxiv.org/abs/1406.5774

[76] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *ISBI*, apr 2015, pp. 294–297.

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

[78] B. v. Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, pp. 286–289.

[79] J. Arevalo, F. A. Gonzlez, R. Ramos-Polln, J. L. Oliveira, and M. A. G. Lopez, "Convolutional neural networks for mammography mass lesion classification," in *Engineering in Medicine and Biology Society*, aug 2015, pp. 797–800.

[80] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 3–24. [Online]. Available: http://dl.acm.org/citation.cfm?id=1566770.1566773

[81] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[82] S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Networks*, vol. 8, no. 78, pp. 1201 – 1214, 1995, automatic Target Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0893608095000615

[83] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, Oct 1996.

[84] D. C. Cirean, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2013, pp. 411–418. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-40763-5_51

[85] ICPR2014. (2014) Mitos dataset. [Online]. Available: https://grand-challenge.org/site/mitos-atypia-14/home/

[86] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, pp. 79–83.

[87] S. Sarraf, G. Tofighi, and for the Alzheimer's Disease Neuroimaging Initiativ, "Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri," McMaster University, Tech. Rep. biorxiv;070441v2, aug 2016. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/070441

[88] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1170–1181, 2016. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7279156

[89] R. Anirudh, J. J. Thiagarajan, T. Bremer, and H. Kim, "Lung nodule detection using 3d convolutional neural networks trained on weakly

labeled data," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2016, pp. 978 532–978 532. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=2507260

[90] z. iek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," *arXiv:1606.06650 [cs]*, Jun. 2016, arXiv: 1606.06650. [Online]. Available: http://arxiv.org/abs/1606.06650

[91] K. Kamnitsas, C. Ledig, V. F. J. 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," *arXiv:1603.05959 [cs]*, mar 2016, arXiv: 1603.05959. [Online]. Available: http://arxiv.org/abs/1603.05959

[92] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *arXiv:1606.04797 [cs]*, Jun. 2016, arXiv: 1606.04797. [Online]. Available: http://arxiv.org/abs/1606.04797

[93] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," *IEEE journal of biomedical and health informatics*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[94] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 652–660. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-24574-4_78

[95] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image deep mining on a very large-scale radiology database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1090–1099. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Shin_Interleaved_TextImage_Deep_2015_CVPR_paper.html

[96] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, Z. Xu, and D. J. Mollura, "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 0, no. 0, pp. 1–6, Jun. 2016. [Online]. Available: http://dx.doi.org/10.1080/21681163.2015.1124249

[97] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[98] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[99] K. Garcia-Reyes, N. M. Passoni, M. L. Palmeri, C. R. Kauffman, K. R. Choudhury, T. J. Polascik, and R. T. Gupta, "Detection of prostate cancer with multiparametric mri (mpmri): effect of dedicated reader education on accuracy and confidence of index and anterior cancer diagnosis," *Abdominal imaging*, vol. 40, no. 1, pp. 134–142, 2015.

[100] 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.

[101] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[102] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[103] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[104] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3361–3368.

[105] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1930–1943, 2013.

[106] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811914005540

[107] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, and ADNI, "Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer #x0027;s Disease," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.

[108] F. Li, L. Tran, K. H. Thung, S. Ji, D. Shen, and J. Li, "A Robust Deep Model for Improved Classification of AD/MCI Patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 5, Sep. 2015.

[109] T. Brosch, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Modeling the Variability in Brain Morphology and Lesion Distribution in Multiple Sclerosis by Deep Learning," in *Medical Image Computing*

and Computer-Assisted Intervention MICCAI 2014, ser. Lecture Notes in Computer Science, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Springer International Publishing, Sep. 2014, no. 8674, pp. 462–469, dOI: 10.1007/978-3-319-10470-6_58. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-10470-6_58

[110] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[111] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 469–477.

[112] L. Zhao and K. Jia, "Multiscale CNNs for Brain Tumor Segmentation and Diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2016, p. e8356294, Mar. 2016. [Online]. Available: https://www.hindawi.com/journals/cmmm/2016/8356294/abs/

[113] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 212–220.

[114] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 115–123.

[115] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative *et al.*, "Latent feature representation with stacked auto-encoder for ad/mci diagnosis," *Brain Structure and Function*, vol. 220, no. 2, pp. 841–859, 2015.

[116] G. van Tulder and M. de Bruijne, "Combining generative and discriminative representation learning for lung ct analysis with convolutional restricted boltzmann machines," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1262–1272, 2016.

[117] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río, "Computer-aided diagnosis of alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences*, vol. 237, pp. 59–72, 2013.

[118] D. Salas-Gonzalez, J. Górriz, J. Ramírez, M. López, I. Alvarez, F. Segovia, R. Chaves, and C. Puntonet, "Computer-aided diagnosis of alzheimer's disease using support vector machines and classification trees," *Physics in Medicine and Biology*, vol. 55, no. 10, p. 2807, 2010.

[119] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[120] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1626–1630.

[121] S. Liao, Y. Gao, A. Oto, and D. Shen, "Representation learning: a unified deep learning framework for automatic prostate mr segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 254–261.

[122] Y. Zhan and D. Shen, "Design efficient support vector machine for fast classification," *Pattern Recognition*, vol. 38, no. 1, pp. 157–161, 2005.

[123] H. R. Roth, J. Yao, L. Lu, J. Stieger, J. E. Burns, and R. M. Summers, "Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications," in *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*. Springer, 2015, pp. 3–12.

[124] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 520–527.

[125] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 246–253.

[126] J. Masci, A. Giusti, D. C. Ciresan, G. Fricout, and J. Schmidhuber, "A fast learning algorithm for image segmentation with max-pooling convolutional networks," *CoRR*, vol. abs/1302.1690, 2013. [Online]. Available: http://arxiv.org/abs/1302.1690

[127] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Control Automation Robotics Vision (ICARCV), 2014 13th International Conference on*, Dec 2014, pp. 844–848.

[128] A. Payan and G. Montana, "Predicting alzheimer's disease a neuroimaging study with 3d convolutional neural networks," in *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, vol. 2, 2015, pp. 355–362, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84938872693&partnerID=40&md5=82d7ac69461e03a425441848043b9d22

[129] H. Roth, C. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. Summers, "Anatomy-specific classification of medical images using deep convolutional nets," in *Proceedings - International Symposium on Biomedical Imaging*, vol. 2015-July, 2015, pp. 101–104, cited By 2. [Online]. Available:

https://www.scopus.com/inward/record.uri?eid=2-s2.0-84944317431&partnerID=40&md5=1e60d74d2b250aa104bf4dde76a3489d

[130] M. Van Grinsven, B. Van Ginneken, C. Hoyng, T. Theelen, and C. Snchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1273–1284, 2016, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84968665432&partnerID=40&md5=d0544991a2092bc95b0dc0e39a1e64fb

[131] A. Bekker, H. Greenspan, and J. Goldberger, "A multi-view deep learning architecture for classification of breast microcalcifications," in *Proceedings - International Symposium on Biomedical Imaging*, vol. 2016-June, 2016, pp. 726–730, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978394324&partnerID=40&md5=de4ad64bcbfc3bed12b78e2683fd2e39

[132] M. Shaken, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, and I. Kokkinos, "Sub-cortical brain structure segmentation using f-cnn's," in *Proceedings - International Symposium on Biomedical Imaging*, vol. 2016-June, 2016, pp. 269–272, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978430327&partnerID=40&md5=bfe78ffe538b5d6c73f7fdfc70019584

[133] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.