Current AI technologies for medical imaging and ethical dilemmas created by them

Zandra Lundegård (35967)

Master of Science in Technology Thesis
Supervisor: Annamari Soini
Department of Information Technology
Faculty of Science and Engineering
Åbo Akademi University
2019

Abstract

Extensive amount of data is collected and documented every day in medical practice, but only a fraction of that data is analyzed and utilized for e.g. diagnosis and treatment plans. Due to the limited capacity of the human brain and lack of time the medical doctors are unable to analyze all the data. Artificial intelligence has the capacity to analyze large datasets.

The aim of this research was to investigate how medical imaging AI could be used in clinical practice and if the technology is reliable and accurate enough to be used in healthcare systems without further development. The ethical concerns of using AI in decision making and diagnosis in medical practice were studied. The results showed that classification, segmentation, and detection built on convolutional neural networks would be a good starting point for implementing artificial intelligence in medical imaging in the future. The findings revealed that the ethical concerns are important to acknowledge and need to be further investigated. Further research within this field could focus on how to develop an ethical framework for AI in medical practice.

Keywords: convolutional neural network, deep learning, radiology

Table of Contents

Acknowledgments

First and foremost, I would like to extend my gratitude to my supervisor Annamari Soini, for keeping me on track, helping, and taking the time to review my work. Additionally, I would like to thank my supervisor Lars Maubach at Accenture, for all the guidance and help with this project.

A special thanks to my family for always supporting me and asking me when I am supposed to graduate. I cannot wait for Christmas, so I finally can tell my grandmother what kind of job my degree will get me. Finally, I would like to thank Mimi for her never-ending support and patience. This project would not have been the same without the help of all parties mentioned.

List of abbreviations and terms

AI	Artificial Intelligence
AUC	Area Under the Curve
ANN	Artificial Neural Network
CBIR	Content-Based Image Retrieval
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
EHR	Electronic Health Record
fCNN	Fully Convolutional Neural Network
GDPR	General Data Protection Regulation
HUS	the Hospital District of Helsinki and Uusimaa
IoU	Intersection over Union
mAP	mean Average Precision
MD	Doctor of Medicine
ML	Machine Learning
SDC	Sørensen-Dice coefficient
SVM	Support Vector Machine

1 Introduction

In modern society, forecasting weather, recognizing faces, detecting fraud, and deciphering genomics are done by artificial intelligence (AI) due to advances in computer science and ultra-fast computing speeds. How AI will impact medical practice is still being thoroughly researched. By processing massive datasets (big data) through layered mathematical models (algorithms), machines learn to detect patterns not decipherable using biostatistics. AI predictive confidence is added through correcting algorithm mistakes, which is referred to as training. AI is being successfully applied for image analysis in radiology, pathology, and dermatology, with diagnostic speed exceeding, and accuracy paralleling, medical experts. It is difficult to reach 100% diagnostic confidence, however combining machines and physicians reliably increases system performance. By applying natural language processing (NLP) to read the rapidly expanding scientific literature and collate years of diverse electronic medical records, cognitive programs are impacting medical practice. Applying AI to medical practice can reduce medical errors, improve subject enrollment into clinical trial, optimize the care trajectory of patients with a chronic disease, and suggest precision therapies for complex illnesses [1].

According to an online survey for hospital staff, answered by staff in several hospital districts in Finland, nearly 33% of doctors spend more than six hours of their workday on a computer. The survey was conducted by the Finnish company Digital Workforce, which offers software for robotic process automation. One out of three of the respondents reported that they spent about thirty minutes every shift typing the same information into several different documents. Almost half of the nurses also reported using more than four hours of every shift on computer-based work. Most of the time was spent on recording and maintaining the patient records [2]. The time-consuming tasks can be completed in minutes with AI [3].

For planning radiotherapy for patients, a doctor studies more than 100 images, each showing a thin slice of the brain. The border of the tumor is carefully marked out by the doctor, image by image. The contours of the sensitive brain region that should be refrained from the radiotherapy beams are also marked out, e.g. the hypothalamus, the pituitary gland and the pathways to the brain's vision centers. The process is very time-consuming and can take hours. The marking needs to be completed before the computers can start calculating how to do the

radiotherapy treatment without harming the important and healthy tissue of the brain [3]. Microsoft has a system called InnerEye [4], that has been tested on prostate cancer patients. InnerEye marks scans automatically, the scans are sent encrypted and anonymized to the InnerEye software. A 3d model of the tumor is created and the information is sent back to the treating doctor. The software was trained on scores of images from previous patients that had been seen and analyzed by experienced consultants, and learned how to mark organs and tumors. The automated process does more than save time, it should perform as well as a top specialist every time, because training is conducted on images marked up by leading experts. The benefit is faster and more precisely delivered treatment [3].

The daily practice of medicine involves repeated situational assessments, pattern recognition dependent on case experience, and evidence-based risk-benefit adjustments. Increasing performance pressure can contribute to relying on information processing shortcuts, e.g. heuristic thinking or gaming cheats, to boost efficiency in decision making and workflow. This might lead to cognitive biases that may foster clinical errors. Human cognitive biases could be avoided by allowing machines to learn directly from medical data and contribute undisputedly to patient care. Humans will have an essential part in the intelligent use of AI in medical practice [1].

Efforts to improve the quality of patient care have increased the use of electronic health records (EHR) and have also introduced a discipline of research utilizing EHR data. Advanced EHRs contain a variety of structured data, e.g. billing and diagnosis codes, electronic prescriptions, and laboratory values. Unstructured data makes up a substantial portion of clinical data, in the form of narrative text notes, either dictated or typed by doctors. A well-known challenge for researchers is extracting correct information from narrative notes which is normally gathered through tiresome medical record review. With the help from NLP, researcher can now extract clinical data from narrative notes in a high flow manner [5].

Approximately 25% of the articles reviewed in the literature review are written by only medical researchers, 50% are conducted as a co-operation between medical and technical researchers, and the remaining 25% of the studies were conducted by researchers with a technical background. The techniques presented in the articles are merely a theory and not in use in clinical practice yet.

1.1 Research questions

The main research question in this thesis is:

1. Is medical imaging AI used in clinical practice?

According to the research question, the main objective of this thesis is to investigate how medical imaging AI is used today in clinical practice. Additionally, to define a research focus, two supporting research questions were defined. The supporting questions aim to reveal the impact that medical classification technology will accomplish in the future and whether the technology is ready to be used in day to day practice without further development.

2. What can be expected in the future?
3. Is the technology accurate enough to be trusted as a reliable technology and used in healthcare systems (without further development)?

The questions are important to gain an understanding of recent AI methodologies, technologies and ethical dilemmas. Answering these questions could be of high interest for a variety of stakeholders. Primary stakeholders could be medical institutions, hospitals, patients and companies developing AI.

1.2 Research scope

This thesis only covers deep learning briefly to gain an understanding of the topic. This thesis will focus on deep learning applications for medical imaging while other applications will be out of scope. Considering medical imaging, the field of radiology will be in scope while other fields, i.e. endoscopy, microscopy, imaging, and visualization, will be scoped out.

1.3 Research purpose

The research aims to evaluate the feasibility of using AI in medical practice using the currently available technologies while investigating the ethics of utilizing automatized diagnostics. Possible implementations were also discussed with interviewees to gain a broader insight into the field.

1.4 Methods

Research approach in this thesis is built on the research ‘onion’ [6], as presented in Figure 1. Starting from the outer layer, first the research philosophy was chosen. Then the research

approach was chosen, the research will be conducted using an inductive research approach. An inductive approach involves exploring the data and developing theories from the data and linking them to the literature. An inductive approach might be involved with the context in which a such phenomenon is occurring. For that reason, studying a small sample of the subject might be more suitable than a larger sample. The research strategy consists of an extensive literature review and interviews. The study is a multi-method qualitative study, meaning that more than one qualitative data collection technique and corresponding qualitative analysis procedure is utilized. The time horizon of the study is cross-sectional, a specific phenomenon is researched at a specific time. Lastly the techniques and procedure were chosen, the data collection techniques are interviews and literature review, and the analysis procedure is qualitative content analysis.

1.4.1 Literature review

In the process, "deep learning" and "deep learning medicine/healthcare" was queried to gain insights on the extent of current literature. There exists a vast amount of literature on deep learning, but slightly less about how to implement deep learning in healthcare solutions. Literature review was then conducted as a systematic review of an extensive part of the literature published on deep learning in health care. Search was conducted via Google Scholar, Arxiv Sanity Preserver and ÅAU's search portal, Alma.

Additionally, the extensive literature review was conducted by posing several questions to the articles and summarizing these in a table, see chapter 5 and Table 5. In total 38 articles were reviewed. The questions posed were which methodology was used, what the purpose of the research was, how quality of the results was assessed, what the background of the researchers was, whether ethical concerns were discussed, whether the research solution is only a theory or whether it is already in use in clinical practice. The questions posed seek to explain how AI and DL are used in medical practice. The question about methodology is important to obtain an overview of the most used methodologies. The purpose of the article is reviewed, to understand the aim of the research. The background of the researchers is reviewed to gain an understanding of whether it is mostly medical researchers, technical researchers, or a co-operation between the two that are conducting the research. The background of the researchers might impact the question whether any ethical concerns are considered. An

important question to pose is if the research is only a theory or if it is already in use in clinical practice, in order to answer the research questions.

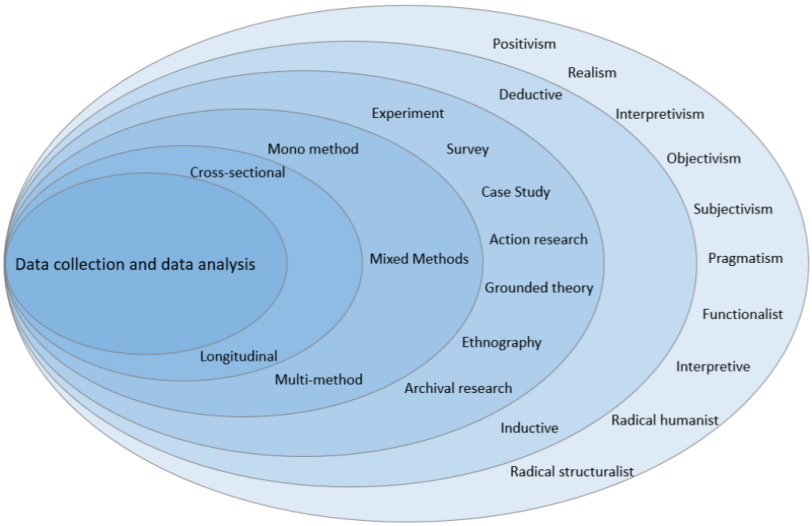

Figure 1 The research approach of this thesis is defined using the research 'onion' [6].

1.4.2 Interviews

The study of this thesis is conducted via semi-structured interviews [6]. The interviews were conducted with experts in the field of AI and doctors familiar with AI. In the interviews, 6 experts were chosen to be interviewed; 3 from Accenture with extensive knowledge of AI and 3 doctors from HUS. Their knowledge about AI in general, possible implementations, and the ethics when using AI in decision-making were researched.

1.5 Limitations and risks

Only one interview was recorded, one interviewee did not consent to recording, and the other four interviews were not recorded due to technical issues. Notes were taken at all interviews and transcribed immediately afterwards. Four interviews were done as Skype calls, which made it more challenging to make proper interviews when the interviewee was not in the same room. Taking notes instead of recording the interviews also interfered with the interviewing.

Limited time and resources interfered with selecting the sample of the study. Ten persons were contacted for interviews, but only six responded. Two companies are represented in the study, which might bias the findings.

1.6 Thesis structure

The methodology of this thesis is shortly described above, a more thorough presentation can be found in chapter 4. In the following sections first, theoretical background about deep learning, with focus on deep neural networks and concepts of medical imaging AI, is presented via literature review in chapter 2. It was considered crucial to reflect on machine learning and especially deep learning, and hence these are first explained to gain extensive insight on the area of deep learning, deep neural networks and deep forests which medical imaging AI is built on. After describing the field of study, a closer look at current technologies in the field is presented in chapter 3. Next the methodology of this thesis is presented in chapter 4, followed by literature review in chapter 5. Results and analysis are discussed in chapter 6. Finally, the thesis is summarized and concluded in the final chapter.

2 Background

A set of computer algorithms that can complete complicated tasks, or tasks that involve intelligence when conducted by humans, is called AI. Machine learning is a subset of AI algorithms which learn and evolve from data and cope without pre-defined rules of reasoning, to complete complicated tasks. Deep learning is a subfield of machine learning problems inspired by the structure and function of the brain called artificial neural networks, which consist of simple interconnected units. An exponential rise in the popularity of deep learning has been enabled due to its capability to process images independently from human intervention. Modifications in position, rotation, scale, perspective, and occlusion can easily be managed. In the medical sector, these features have consequently appeared to valuable. The amount of analyzable image data is constantly increasing, due to the modernization and constant use of imaging devices. Before the arrival of deep learning, time-saving decision support was attained by machine learning techniques, but with a different supplementary human cost, i.e. labeling and highlighting the interesting regions in every image was done manually. Deep learning provides accurate results for image processing and image interpretation. Deep learning makes expertly handcrafted features redundant by automatically learning the optimal attributes from the available images, and improving due to large amounts of available data. Deep learning can be applied on many tasks, such as landmark detection, tissue segmentation, diagnosis, and prognosis.

2.1 What is medical imaging?

Medical imaging includes techniques and processes designed to visualize body parts, tissue or organs for medical purposes, such as diagnostic and treatment purposes. It covers disciplines such as radiology, endoscopy, microscopy, imaging, and visualization. To narrow the scope of this thesis, it will focus on radiology, as there exist many articles and extensive literature on that field. Radiology means imaging the inside of the human body for the purpose of making a diagnosis. The term medical radiology includes diagnostic radiology as well as intervention, i.e. treatment guided by images.

Radiology is the branch of medicine that concentrates on using medical images for detection, diagnosis and characterization of disease (diagnostic radiology) as well as guiding procedural interventions (interventional radiology). Medical image interpretation is also present in other

branches of medicine including cardiology, orthopedics, and surgery. Personal interaction is often poor between radiologists and patients, and other physicians in diagnostic radiology. To view an image and generate a written report of the findings is the primary task of a radiologist. The well-structured and isolated nature of the radiologist's work makes it exceptionally attractive as application domain for AI algorithms. AI algorithms in general and deep learning techniques have an incredible potential to shape the practice of radiology. Nearly all the primary data handled in imaging is digital, and amenable to analysis by AI algorithms [7].

Development over the last few years of CT, MRI, ultrasound, and software for three-dimensional imaging offers entirely new opportunities for diagnosis and treatment. At the molecular and cellular level, the new techniques will make diagnosis possible before symptoms appear, and they allow individually-adapted gene-based therapy with great precision.

2.2 Image processing

Image processing is the use of quantitative analyses and/or algorithms to perform processing on digital images. It enables generation of 3D parametric maps and involves calculation of values that should be ultimately replicable and rater-independent. Image processing methods are developing rapidly, and the trend is to integrate as much automation as possible [8]. Tasks that are crucial for this thesis are classification, object detection, and instance segmentation.

The purpose of image classification is to map a given image to a limited set of class labels, determining which one of a set of categories an image belongs to. In other words, image classification is a mapping process from one vector representation, the source image, to another vector representation, the output vector. This is quite a challenging task to solve for a computer whereas a human can take one look at an image and classify the elements. A classic image classification problem is to determine if an input image contains a cat or a dog [9]. The computer learns to do this by training on images known to contain a cat and images known to contain a dog. A trivial task for a human, but a computer needs an algorithm to learn from the known images, and then uses the gained knowledge to classify new input images as a cat or a dog.

Image classification models classify images into a single category, usually corresponding to the most distinct object. Images are usually complex and contain multiple objects. Assigning a

label with image classification models is complicated and uncertain. The purpose of object detection is to produce a set of tight boxes around objects, in the given image, while automatically classifying them. Several labels can be assigned to one image in object detection. The purpose of instance segmentation is to solve detection and segmentation jointly. Instance segmentation has attracted a considerable amount of attention. Its potential applicability to a wide area of applications and the stimulating technical challenges are the motivation for the interest. Instance segmentation is more challenging than other pixel-level problems. The number of groups (instances) is unknown a priori in instance segmentation, whereas in semantic segmentation, which deals with classifying every pixel of an image, each pixel belongs to a set of predefined groups [10]. Figure 2 illustrates the differences between classification, object detection and instance segmentation.

Image classification can be implemented to classify tissue as either normal or tumor tissue in radiological images. Object detection can be applied to applications to highlight interesting areas in radiological images for physicians to analyze further. Instance segmentation can be used to build 3D models of tumors and healthy anatomy in radiological images.

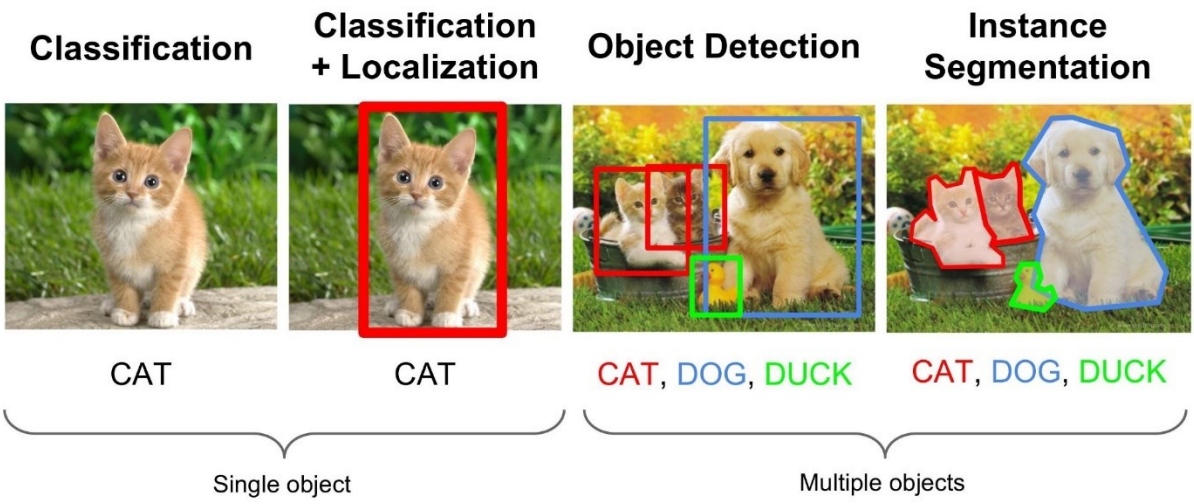

Figure 2 Comparison between image classification, object detection and instance segmentation [11].

2.3 Deep learning

2.3.1 Terminology

It is crucial to first grasp the related concepts of AI and machine learning in order to understand deep learning and medical imaging AI. Deep learning is a subfield of machine learning problems inspired by the structure and function of the brain called artificial neural

networks, which consist of simple interconnected units. The units are connected and form multiple layers that can generate increasingly high-level representations of the provided input, e.g. images. In Figure 3, the relation between artificial intelligence, machine learning, and deep learning is visualized.

Below, artificial neural network is explained in general, and one specific type of deep neural network, convolutional neural network, is introduced more thoroughly to explain the architecture of deep learning models. The learning process of these networks, which is the process of integrating the patterns obtained from data into deep neural networks, is then explained in detail [7].

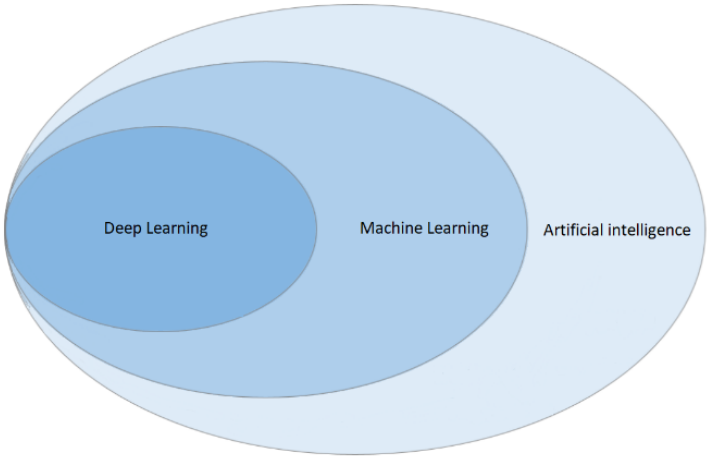

Figure 3 The relation between AI, ML, and DL.

2.3.2 Artificial neural networks

One of the main models used in machine learning is artificial neural networks (ANNs). An ANN is inspired by the human brain and is a mathematical representation of the human neural architecture intended to replicate the way humans learn. An ANN is a system consisting of several parallelly connected, computationally simple but nonlinear elements that can process information from the environment. Due to its large network, massive parallelism, optimal spatial organization, associative memory, and ability to reason by analogy, the human brain often outperforms ANNs in terms of processing speed, even though the response of the neurons in the human brain is slow in comparison to ANNs.

Artificial neurons are called units. A typical ANN consists of anything from a few dozen to hundreds, thousands, or even millions of units arranged in layers. Each layer connects to the layers on both sides. Input units receive information in various forms from the outside world,

and the network attempts to learn about, recognize, or process the information. Output units are on the opposite side of the network and signal how it responds to the information it has learned. One or several layers of hidden units are in between input units and output units. If each hidden unit and each output unit is connected to every unit in the layers either side, the ANN is fully connected, see Figure 4. A weight represents the connections between one unit and another. The weight can be either positive or negative. The influence one unit has on another is dependent on the weight; a higher weight has more influence than a lower weight [12].

The interconnected nature of the network supports the performance of highly sophisticated calculations and implementation of complicated functions, even though every individual neuron performs only a simple calculation [7]. Due to their ability to model highly non-linear systems in which the relation among the variables is unknown or complex, ANNs are extensively applied in research [13].

Information flows through ANNs in two ways. A common design for ANNs is called a feedforward network. Information is fed into the network via the input units when the network is being trained or operating normally, which triggers the layers of hidden units, and information arrives at the output units. The units to the left deliver inputs to the units to their right, and the inputs are multiplied by the weights of the connections they travel along. In the simplest type of network, every unit adds up all the inputs it receives and if the sum is more than a certain threshold value, the unit “fires” and triggers the units it is connected to. ANNs learn in the same way, normally by a feedback process called backpropagation. Backpropagation is done by comparing the output the ANN produces with the intended output and using the difference between the two outputs to adjust the weights, going backward from the output units through the hidden units to the input units. Backpropagation causes the ANN to learn over time, minimizing the difference between actual and intended output until they coincide, reducing wrongful predictions [12].

2.3.3 Convolutional neural networks

Deep neural networks (DNNs) are a combination of deep learning and neural networks, and a special type of ANNs. DNNs are a crucial part of deep learning [14]. Classic DNNs are fully connected, neurons have full connections to all activations in the previous layer. Neurons connect across adjacent layers, never within a layer. A deep convolutional neural network

(CNN) is the most common type of a DNN [7]. Convolution is an operation, whose main goal is to extract features from the input image [15].

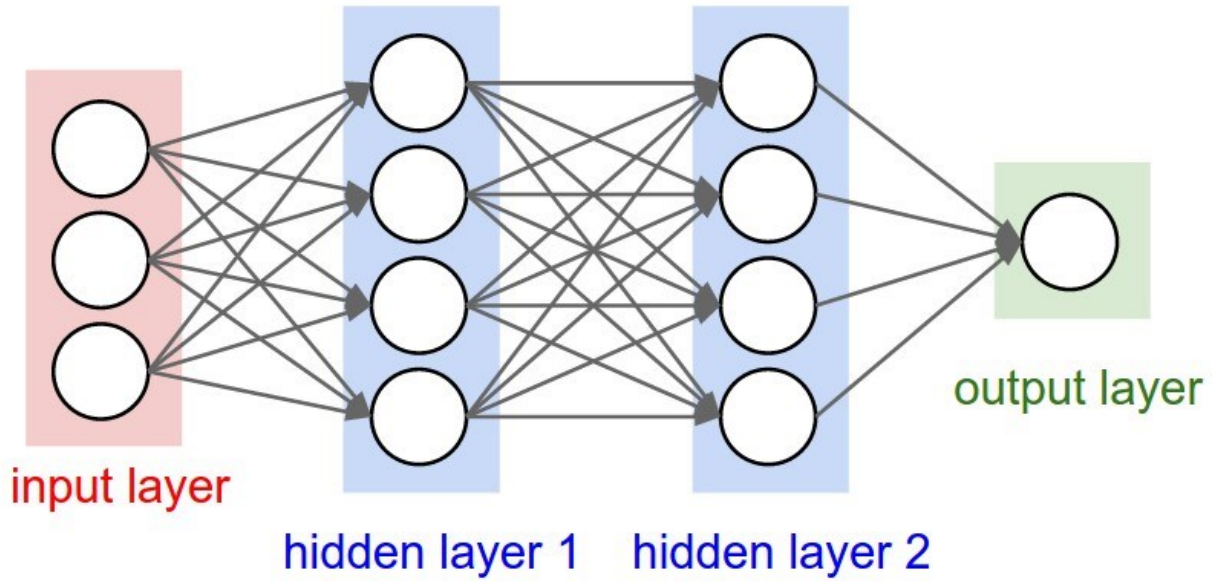

Figure 4 A fully connected ANN [16].

Convolutional neural networks (CNN) are the main approach for image analysis. The relevant features are detected gradually, due to the special convolutional layers, from low-level to high-level structures, by inspecting small fractions of the training images (this is the feature discovery). As a result of the independent feature discovery, the image is labeled according to the given task [17]. CNN modeling may target images from histopathology, computed tomography (CT) scans, or magnetic resonance imaging (MRI). Automatic feature extraction can be done by stacked autoencoders, such as in [18] for MRI imaging and in [19] for CT and ultrasound images. The topic of deep learning for medical imaging is constantly present at conferences specialized in biomedical computing and in journals dealing with medical image processing. Despite the many benefits of medical image processing, some disadvantages when applying it to real-world complex images, as compared to its application to general images, must be considered. The lack of labeled medical images interferes with the performance, contributes to overfitting and hard parametrization. Overfitting refers to a model that models the data too well. Noise or random fluctuations in the training data are learned as concepts by the model. These concepts do not relate to the new data and decrease the model's ability to generalize. Data augmentation, transfer learning, and fine-tuning from general data sets

such as ImageNet or those on Kaggle, are current solutions that cope with the disadvantages [17]. Data augmentation is the process of artificially generating training data through different ways of processing, such as random flips, rotation, and shifts [20]. In transfer learning, already gained knowledge from solving one problem is applied to solve a different but related problem. The network is first trained using a different dataset. ImageNet [21] is an image database, created to provide researchers an easily accessible image database. For example, an ImageNet collection is used to train a network. Then additional training, with data specific to the problem, is done to fine-tune the network. A certain level of processing, such as recognition of edges or simple shapes, can be shared when solving different visual tasks. Fine-tuning refers to adjusting weights in order to achieve better or a desired performance.

Several factors have made it possible for deep learning to thrive [22]. The availability of large datasets is a key factor for the success of deep learning. The performance of DNNs strongly rely on the capacity of the model, which can be improved by adopting deep and wide architectures. The numbers of parameters increase with the improved network capacity, and therefore require more data to reliably estimate them. Fortunately, the extensive access to internet and smartphones favors easy and cheap big data collections. Extensive computational power is required to accordingly exploit deep models and large datasets. Development of specialized hardware for deep learning has been in focus in the last years. Most deep learning practitioners use modern GPUs to efficiently train complex models. Considerable computational power is needed by the DNNs during the training phase.

A matrix of pixel values can represent an image. Channel refers to a certain component of an image. An image from a conventional camera has three channels; red, green and blue (RGB). The three channels can be imagined as three 2d-matrices stacked on top of each other, with pixel values in range 0-255. A grayscale image has only one channel [15].

There are four main operations in every CNN, these are convolution, non-linearity (ReLU), pooling and classification (fully connected layer). The pooling operator reduces the number of parameters when the image is too large. The convolution operator in a CNN extracts features from the input image. Table 1 represents an image whose pixel values are only 0 or 1 (a special case of a grayscale image, normally pixel values for grayscale images range from 0 to 255). A convolution operation extracts features from the input image. The spatial relationship

between the pixels is preserved when the convolutional operation is performed, by learning image features using small squares of input data, see Figure 5 [15].

Table 1 Left: a simple black and white 5x5 image, represented as a 5x5 matrix. Right: A 3x3 image filter [15]

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

1	0	1
0	1	0
1	0	1

Figure 5 shows how the convolution of the 5x5 image and the 3x3 matrix can be computed. The 3x3 matrix is called a filter or kernel or feature detector. A convolution operation is performed by sliding the 3x3 image filter over the 5x5 image and multiplying the 3x3 area of the image that is covered by the kernel. This results in a 3x3 result matrix. The result matrix represents the degree of overlap between the image and the kernel. The convolved feature or activation map or feature map is the matrix formed by sliding the filter over the image and computing the dot product. Filters act as feature detectors from the original input image.

Rectified Linear Unit (ReLU) is a non-linear operation. ReLU replaces all negative pixel values in the feature map by zero and is an elementwise operation applied per pixel. Most of the real-world data the CNN needs to learn is non-linear. To account for non-linearity ReLU is introduced, since convolution is a linear operation.

Dimensionality of each feature map is reduced with spatial pooling, while retaining the most important information. Spatial pooling is performed between convolutional layers and fully-connected layers, with the aim to map any size input down to a fixed size output. There are different types of pooling, e.g. max, average, and sum. Max pooling takes the largest element from the rectified feature map within a spatial neighborhood, e.g. a 2x2 window. Average pooling takes the average of all elements in the defined window and sum pooling sums all elements [15].

CNN architectures explicitly assume that the inputs are images, which allows encoding certain properties into the architecture, such as if some feature is beneficial to compute at some spatial position (x_1, y_1), then is should also be beneficial to compute at a different position (x_2, y_2) [16]. The forward function is more efficient to implement and the number of

parameters in the network is vastly reduced, due to this assumption. A parameter is a property of the training data learnt during training, e.g. weights of the connections.

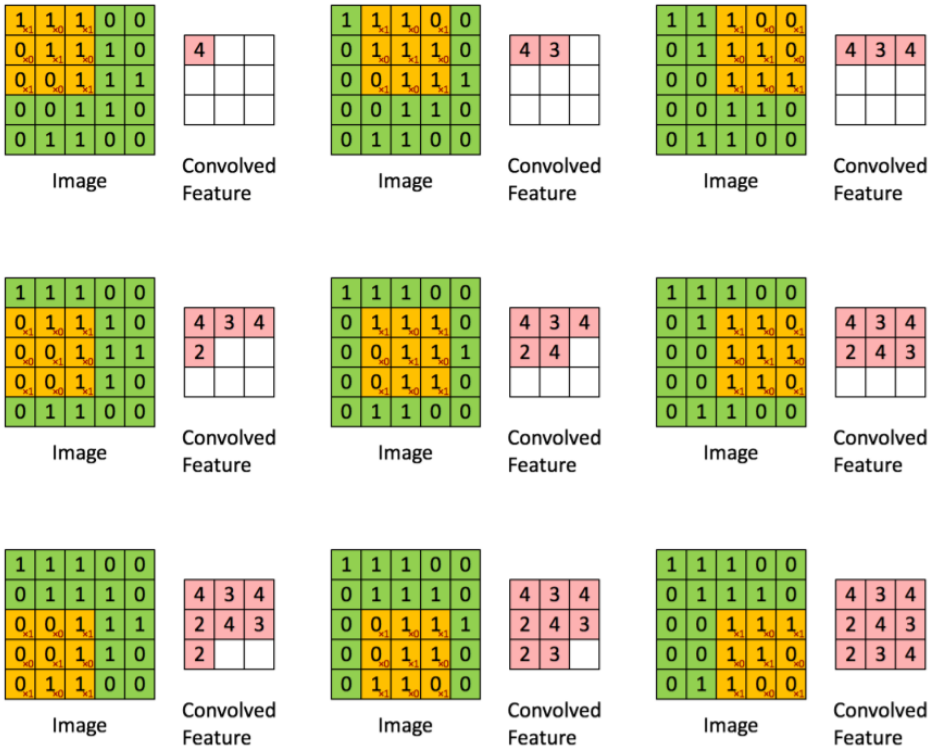

Figure 5 Example of a convolution operation [15].

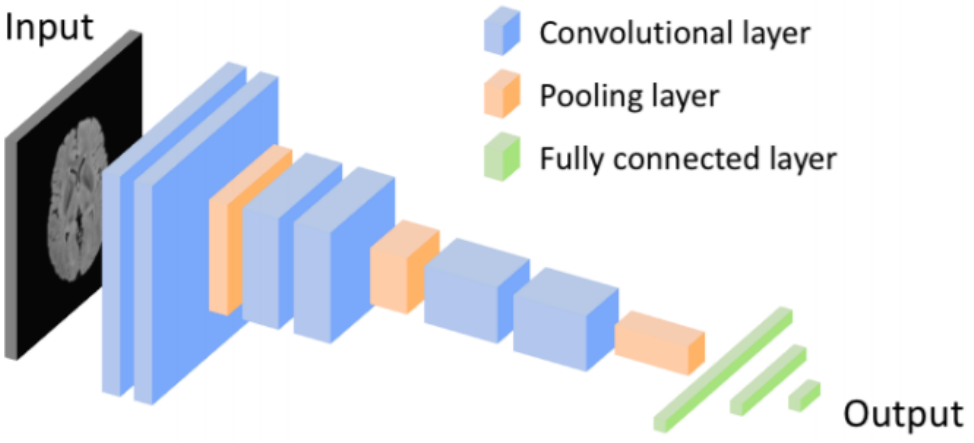

Figure 6 A typical architecture of a convolutional neural network [7].

An example of a small architecture of a CNN is presented in Figure 6. The convolutional layers, which are the first ones, generate useful features for classification. A filter is a small matrix used to apply effects on images, such as blurring, sharpening, outlining, or embossing, in order

to represent the image in a different way. The first convolutional filters can be considered as implementing image filters, varying from simple filters that match edges to those that eventually match highly complicated shapes such as eyes. The fully connected layers utilize the features extracted by the convolutional layers to produce a decision, e.g. assign a label to an image. Driven by the characteristics of the task at hand, a variety of deep learning architectures have been proposed, e.g. fully convolutional neural networks for image segmentation [7].

2.3.4 The learning process in convolutional neural networks

The networks are taught to perform useful tasks in the process referred to as learning or training. Three types of learning processes exist, i.e. supervised, semi-supervised, and unsupervised learning. The most popular is supervised learning, where all training data is labeled, and the algorithms learn to predict the output from the input data. For example, the network takes an image (e.g. image of a tumor) as input and calculation is done within the network to produce a prediction (e.g. if the tumor is benign or malignant) based on the current weights of the network. An error is calculated by comparing the predication to the actual label of the image. To adjust the values of the network's weights this error is propagated through the network (backpropagation) and the next time the network analyzes this example, the error decreases. In practice, after a group of examples are presented to the network the adjustment of the weights is performed. The weights are adjusted in the direction where the error decreases, this is an iterative process that consists of calculating the error between the output of the model and the desired output.

To start with a random set of weights and train the network using available data specific to the problem being solved is referred to as training from scratch and is the most straightforward way of training. A limited amount of training data is common in medical imaging and with the large number of parameters, often above 10 million, a network may overfit the available data, resulting in low performance on the test data. Transfer learning and off-the-shelf learning (also referred to as deep features) have been developed to cope with this issue. Figure 7 visualizes the difference between training from scratch with transfer learning and off-the-shelf deep features [7].

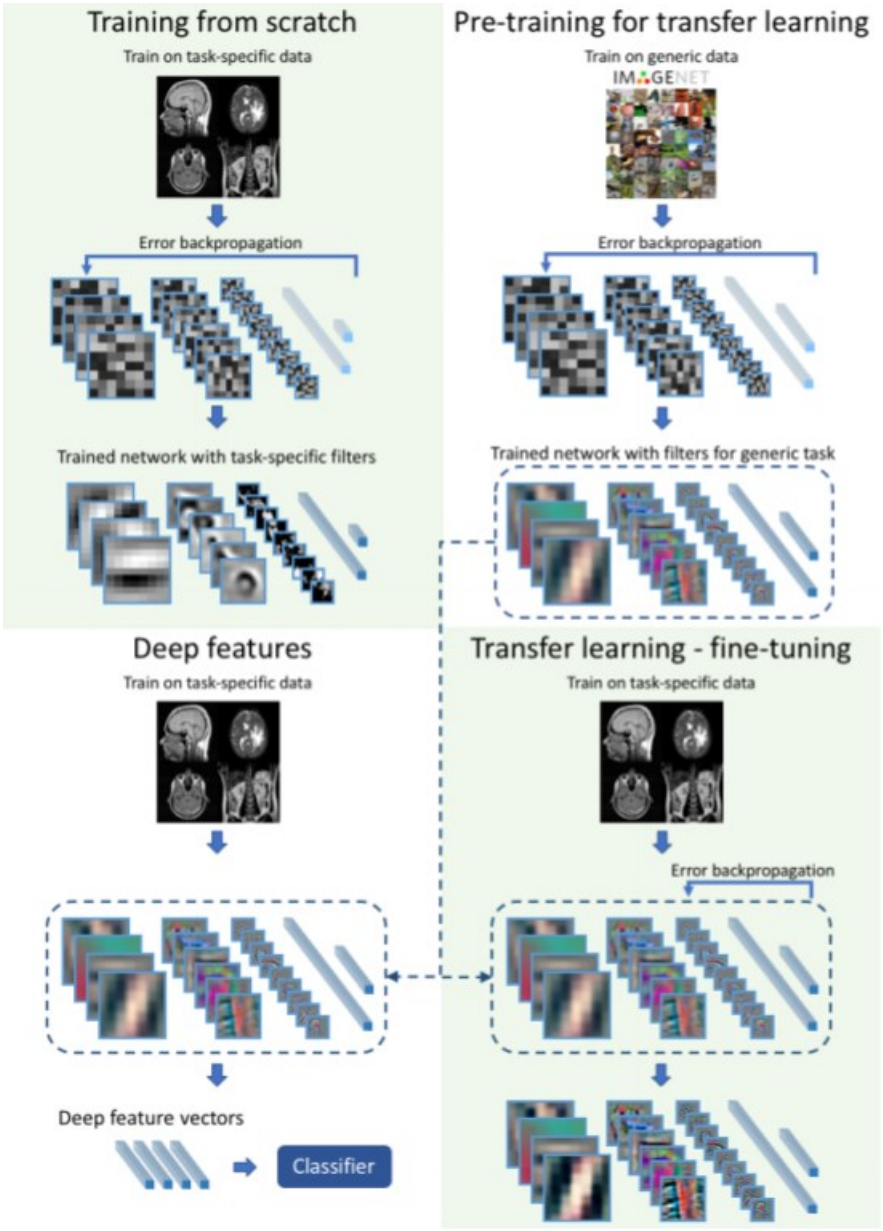

Figure 7 Different ways of training a deep neural network [7].

Prediction of survival time from brain MRI in patients with glioblastoma tumor [23] or in skin lesion classification [24] has successfully benefited from transfer learning. The issue of limited training data can also be handled with a deep off-the-shelf features approach. This approach uses pretrained CNNs, which have been trained on a different dataset, e.g. an ImageNet collection, to extract features from the images by extracting outputs of layers prior to the network's final layer. Those layers generally consist of hundreds or thousands of outputs. The outputs are then fed to "traditional" classifiers, such as linear discriminant, support vector machines, or decision trees as inputs. This approach has some similarities with transfer learning, and is sometimes viewed as part of transfer learning, but the last layers of a CNN are

changed to a traditional classifier and no additional training for the early layers is done [7]. Linear discriminant is a method which attempts to discover a linear projection of high-dimensional observations into a lower-dimensional space [25]. Support vector machine (SVM) is a supervised machine learning algorithm. In SVM data items are plotted as a point in n-dimensional space, the value of a coordinate is the value of each feature [26]. Decision trees are utilized to visually and explicitly represent decisions and decision making, this method uses a tree-like model of decisions [27]. The dataset is broken into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The decision tree consists of one root node with several decision and leaf nodes, see Figure 8.

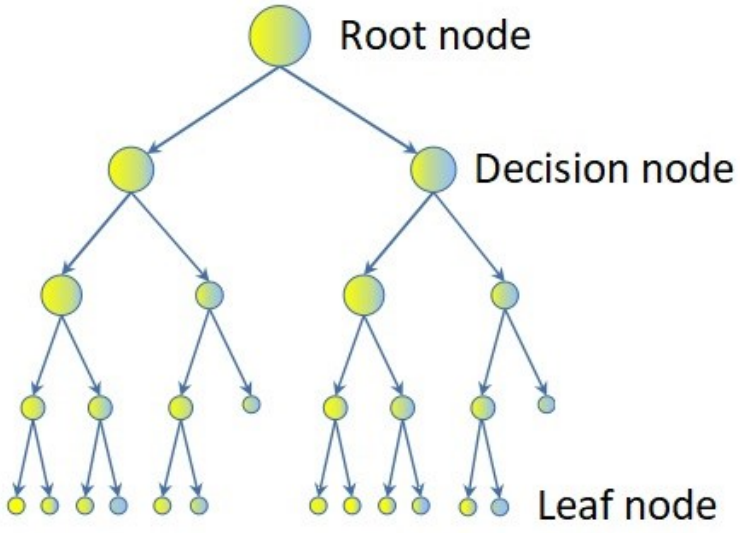

Figure 8 Visualization of a decision tree [28].

2.3.5 Deep neural decision forests

Most of the current deep learning models are built upon neural networks. Deep neural decision forest is a novel approach, presented by [29], that merges decision trees and the representation learning functionality known from deep CNNs, by training them in an end-to-end manner. In end-to-end training, intermediary algorithms are excluded and the solution to a given problem from the sampled dataset is directly learnt. Combining decision trees with convolutional networks forms deep neural decision forests, which is a stochastic and differentiable decision tree model, which drives the representation learning, generally handled in the initial layers of a convolutional network. Differentiable refers to the computation of the derivate of the operations in the module, for example when utilizing

backpropagation, the gradient of the loss function is computed with respect to the module parameters. Representation learning attempts to learn representations of the data, making it easier to extract valuable information when building classifiers or other predictors [30]. In [29], the model differs from both conventional decision forests and DNN, because a decision forest provides the final predictions. This model is employed in the research project InnerEye conducted by Microsoft [4].

[31] propose a deep forest model, the gcForest, based on nondifferentiable modules. The model generates deep forests with layer-by-layer processing, in-model feature transformation, and sufficient complexity, and is a decision tree ensemble approach with fewer hyper-parameters than DNNs. In-model feature transformation performs a linear combination of the original features, e.g. pre-processing data to scale several features to a common value range to make the contribution from all features equal. The hyper-parameters are the variables that determine the network structure, e.g. number of hidden units. The value of hyper-parameters is set before the learning process begins. The research demonstrates the possibility of designing deep models based on non-differentiable modules and without utilizing backpropagation.

2.3.6 Deep learning vs 'traditional' machine learning

There is a distinction between deep learning and traditional machine learning, see Figure 9. Especially in the context of medical imaging the difference is extremely important. Feature extraction is the regular first step in traditional machine learning. One must determine which characteristics of an object will be significant and implement algorithms that can capture these characteristics to classify an object. In the field of computer vision, several sophisticated algorithms have been introduced for this purpose and a variety of size, shape, texture and other features have been extracted. This process is arbitrary. Often the machine learning researcher or practitioner must predict which features will be useful for a specific task and (s)he faces the risk of including useless and redundant features and excluding useful features. No decisions regarding which features should be extracted need to be made in deep learning, because the processes of feature extraction and decision making are merged and trainable. For much larger training data sets, the cost of allowing the neural network to choose its own features is a requirement [7].

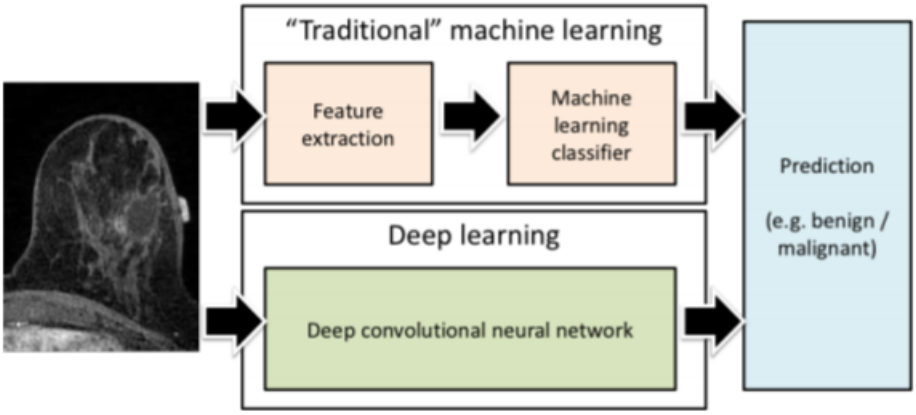

Figure 9 Difference between traditional machine learning and deep learning illustrated [7].

2.4 Why the need for deep learning in medical practice?

All fields of medicine, from drug discovery to clinical decision making, have the potential to extensively apply deep learning algorithms, significantly altering the way medicine is practiced. An increasing amount of medical records are digitalized and the success of deep learning algorithms at computer vision tasks in recent years comes at an opportune time. Amongst office-based physicians in the US, the use of EHR increased from 11.8% to 39.6% during 2007-2012 [32]. Analyzing medical images is a crucial part of a radiologist's work, where a human is limited by speed, fatigue, and lack of experience. To train a qualified radiologist takes years and great expense, and some health-care systems outsource radiology reporting to lower-cost countries such as India via tele-radiology. An incorrect and delayed diagnosis can be fatal for the patient. Thus, an automated, accurate and efficient deep learning algorithm is ideal to carry out medical image analysis with. The data in medical image analysis is relatively structured and labeled, and this is partly the reason why it is an active field of research in deep learning. Radiology is likely to be the area where patients first encounter functioning and practical AI systems. This is important for two reasons. Firstly, medical image analysis is an indicative test as to whether AI systems can advance patient outcomes and survival. Patient outcome refers to the changes in health that occur from measures or specific health care investments or interventions [33]**.** Secondly, it serves as a testbed for human-AI interaction, of how responsive patients will be towards health altering choices being made or assisted by a non-human actor [34].

Rapid differentiation of abnormalities from normal background anatomy is one of the most demanding tasks in the interpretation of imaging. To detect a small number of suspicious or

abnormal findings, each radiograph contains thousands of individual focal densities, regional densities, and geometric points and lines that must be interpreted. The task grows even more complex when the entire mammogram is required to be interpreted as either normal or negative. A computer algorithm is not required to detect all objects of interest (e.g. abnormalities) and be completely specific to be useful.

The often complex task of determining a diagnosis and the disease management implications is initiated once an abnormality has been identified. To decide how to accordingly manage the finding, many features must be integrated for focal masses generically. These features contain size, location, attenuation or signal intensity, borders, heterogeneity, and change over time, to name a few. Making the diagnostic management decision follows straightforward guidelines for some types of abnormalities, while for other management the algorithms are very complex. The radiologist determines whether a mass is likely to be benign or requires follow-up or biopsy, based on the constellation of features. Many features can be assessed potentially with deep learning algorithms, even those previously dismissed by radiologists, and reach a repeatable conclusion (a repeatable conclusion is the same conclusion that a human would arrive at under the same conditions) in a fraction of the time required for a human interpreter. A process that currently is extremely time-consuming, laborious, and one that requires human interpretation, is to categorize large amounts of existing imaging data and correlate features with downstream health outcomes (i.e. disease, injury, and mortality). This could be solved using deep learning algorithms [7].

3 Deep learning in radiology

In this section, an overview of applications of deep learning in radiology is given. This section is organized by the tasks that the deep learning algorithms perform. Within each subsection, different methods applied are described, and when possible, the evolution of these methods in the recent years is discussed.

3.1 Classification

One of the first areas in which deep learning made a significant improvement to medical analysis was image or exam classification. Typically, in exam classification, one or multiple images (an exam) are fed to the network as input and a single diagnostic variable is returned as output (e.g. disease present or not) [35]. In radiology, several different classification tasks occur such as: classification of an image or an exam to determine abnormality present or not; classification of abnormalities as benign or malignant; classification of cancerous lesions according to their histopathological and genomic features; prognostication; and classification for organization of radiological data [7].

For classifying radiological data, deep learning is becoming the methodology of choice. CNNs with a varying number of convolutional layers followed by fully connected layers are used by most available deep learning classifiers. Compared to the natural image datasets, which have pushed the development of deep learning techniques in the last five years, the availability of radiological data is limited. Dataset sizes in exam classification are generally hundreds/thousands of samples versus millions of samples in computer vision [35]. Off-the-shelf features and transfer learning ease this issue, and many deep learning applications in medical image classification have applied these techniques. In a variety of domains, off-the-shelf features have performed well [36], and have been successfully applied to medical imaging [37, 38]. The deep off-the-shelf features extracted from a pre-trained VGG19 network and hand-crafted features for determining malignancy of breast lesions in mammography, ultrasound, and MRI were combined by the authors in [37]. Prediction of long-term and short-term survival was conducted in [38] for patients with lung carcinoma. Transfer learning has been applied to tasks such as classification of prostate MR images to distinguish patients with prostate cancer from patients with benign prostate conditions [39], and identification of CT images with pulmonary tuberculosis [40]. Shallow layers are fixed after the initial training and

the deepest layers are replaced and retrained in most of the studies which apply transfer learning strategy. Combining fine-tuning and deep features approach is a variation of the transfer learning strategy. To achieve more task-specific deep feature representations, a pre-trained network is fine-tuned on a new dataset. An example of this is the study [41], which used features extracted from a fine-tuned pre-trained GoogLeNet to perform ultrasound imaging-based thyroid nodule classification. In the study [42] a collection of fine-tuned CNN classifiers was presented to predict radiological image modality. Approaches using deep features and transfer learning with fine-tuning were compered in [43] and [44]. Deep features performed better than transfer learning with the fine-tuning approach in both problems. A small training dataset was an issue faced in both studies.

A complete DNN can be trained from scratch when the size of the dataset is sufficient. The task and dataset characteristics determine the size of the network to be trained. Modifications of AlexNet [45] and VGG [46], with fewer layers and weights, are the generally used architecture in medical imaging. Various studies demonstrate training from scratch, such as assessing for the presence of Alzheimer's disease based on brain MRI using deep learning [47, 48], glioma grading in MRI [49], and disease staging and prognosis in chest CT of smokers [50]. Networks are easier to train and more efficient, due to recent advances in the design of CNN architectures. Having fewer trainable parameters reduces the probability of overtraining [51], while the network has more layers and performs better. Applications of deep learning for radiology have also shifted to these more powerful networks both for transfer learning and training from scratch.

Radiological text report analysis is also an important task in radiology, apart from classification of radiological images. Deep learning-based natural language processing (NLP) is the most notable approach in this type of classification, which is established on the seminal work for gaining vector representation of phrases applying an unsupervised neural model. This architecture is illustrated in [52], where the authors label CT radiology reports as displaying presence or absence of pulmonary embolism, including type (chronic or acute) and location of pulmonary embolism when present.

3.2 Segmentation

Quantitative analysis of clinical parameters related to volume and shape, for example cardiac or brain analysis, is enabled by segmentation of organs and other substructures in medical images. Segmentation is a crucial first step in computer-aided detection [35]. In both natural and medical image analysis, segmentation is a frequent task. An image is divided into different regions to classify each pixel in the image with CNNs, by presenting it with patches extracted around the pixel. The common applications are segmentation of organs, substructures or lesions in radiology, generally as a preprocessing step for feature extraction and classification [7].

Classification of individual pixels, based on small image patches (both 2- and 3-dimensional) extracted around the classified pixel, is still the most uncomplicated and widely used method for image segmentation. The main issue of this naive 'sliding-window' approach is computationally inefficient since input patches have overlapping parts of the image and the same convolutions are computed multiple times. The inner products can be written as convolutions and vice versa, due to the convolution and dot product being both linear operators. The CNN can take larger input images than it was trained on and produce a likelihood map, by rewriting the fully connected layers as convolutions [35]. Each pixel is segmented based on a limited-size context window and overlooks the wider context. For example, pixel location or relative position to other image parts, i.e. a piece of global information, may be required to accurately assign the label of the pixel in some cases.

Fully convolutional neural network (fCNN) [53] is one approach that deals with the issue of pixel-based segmentation. The entire image (or a large part of it) is processed by the fCNN at the same time and outputs a 2-dimensional segmentation map instead of a label for a single pixel. In radiology fCNNs have been applied to several tasks [54, 55, 56]. Class imbalance is typical in medical datasets, i.e. the number of examples differs in every class. Loss functions have been researched to address the imbalance. The loss function represents the price paid for inaccuracy in predictions in classification problems.

In terms of trainable parameters, 3-dimensional fCNNs are significantly larger and require significantly larger amounts of data. It is common to process data as 2-dimensional slices and then merge the 2-dimensional segmentation maps into a 3-dimensional map, to segment 3-

dimensional data. There are successful applications of 3-dimensional fCNNs in radiology, where these obstacles are overcome, e.g. prostate segmentation from MRI with V-Net [57], segmentation of the proximal femur in MRI [58] with 3D U-Net [59] and tumor segmentation [60]. Recurrent neural networks (RNNs) is another deep learning approach that has been applied to some medical imaging segmentation tasks [7].

3.3 Detection

A crucial pre-processing step in segmentation tasks or in the clinical workflow, for therapy planning and intervention, is anatomical object localization (in space or time), such as organs or landmarks. Parsing of 3D volumes is often required in medical imaging localization. Several approaches have been proposed to solve 3D data parsing with deep learning algorithms, e.g. treat the 3D space as composition of 2D orthogonal planes. The most popular strategy overall with good results to identify organs, regions and landmarks, seems to be localization through 2D image classification with CNNs. According to [35], research is conducted on this concept with emphasis on accurate localization by modifying the learning process. The authors predict such strategies to be researched further to show that deep learning techniques can be modified to an extensive range of localization tasks, for example multiple landmarks.

One of the most labor-intensive key parts of diagnosis for clinicians is the detection of objects of interest or lesions in images. Localization and identification of small lesions in the full image space is typically part of the task [35]. Computer-aided systems that automatically detect lesions have a long research tradition. Already in 1995, the first object detection system applying CNNs was proposed [61]. The CNN had four layers and was designed to detect nodules in x-ray images.

A 2-phase process is the most common approach to detection for 2D data that requires training of two models. All suspicious regions that may include the object of region are identified in the first phase. This phase requires high sensitivity and it normally generates many false positives. A regression network for bounding box coordinates is a common deep learning approach for the first phase, based on classification architectures [62, 63]. The second and last step is to classify the sub-images extracted in the previous step. Some applications only apply deep learning in one of the two steps. When applying deep learning in the second step, transfer learning is usually applied to perform the classification. Pre-training on other

medical imaging datasets is utilized in other applications, such as in [64] to detect masses of 3D mammography images. The network architectures used in a regular classification task can be utilized for the second phase, e.g. VGG [46], GoogLeNet [65], Inception [66], and ResNet [67].

In the end-to-end approach one model, including both phases, is trained, while the models are trained separately for each phase in the 2-phase detection process. The faster region-based convolutional neural network (R-CNN) [68] is an end-to-end architecture that has proven success in object detection in natural images. The R-CNN was recently applied to medical imaging. A feature map is obtained by utilizing a CNN. The feature map is shared between a region proposal network that outputs bounding box candidates and a classification network. Each category is predicted by the classification network [7].

3.4 Other tasks

While classification, segmentation, and detection are the main tasks to solve for deep learning in radiology applications, other medical imaging-related tasks have also benefitted from deep learning. No unifying methodological framework exists yet for these solutions, due to the variety of these problems. The examples below are arranged according to the problem that they attempt to solve.

3.4.1 Image registration

A common image analysis task is registration, i.e. spatial alignment, in which a coordinate transform is calculated from one medical image to another. For example, two or more images typically of different types (e.g., T1-weighted and T2-weighted MRIs) need to be spatially aligned so that the same location in each image displays the same physical location in the pictured organ [7]. This task is often done in an iterative manner where a parametric transformation is presumed and a pre-determined metric (e.g. L2-norm) is optimized. Although registration is not as popular as segmentation and lesion detection for deep learning, getting the best possible registration performance can benefit from DNNs. Utilizing deep learning networks to estimate a similarity measure for two images to drive an iterative optimization strategy, and to directly predict transformation parameters using deep regression networks are two strategies that are popular in current literature.

The best way to integrate deep learning techniques in registration methods is not yet settled in the research community. Only a few papers exist on the subject and these have distinctly different approaches. [35] expect more contributions of deep learning in medical image registration soon.

3.4.2 Image generation and enhancement

Techniques ranging from removing obstructing elements in images, normalizing images, improving image quality, and data completion to pattern discovery are some of the image generation and enhancement methods using deep learning that have been proposed. 2D and 3D CNNs are utilized to convert one input into another in image generation. The pooling layers, present in a classification network, are typically not present in these architectures. The input and the desired output are present during the training of the system. The loss function is defined as the differences between the generated and desired output. Creative applications of deep learning in significantly differing tasks have enabled impressive results for image generation. These tasks are expected to increase further in the future [35].

Image enhancement focuses on developing different characteristics of the image such as resolution, signal-to-noise-ratio, and necessary anatomical structures (by suppressing unnecessary information) through various approaches such as super-resolution and denoising. Especially in cardiac and lung imaging, super-resolution of images is crucial. Long scan times are often required in 3D near-isotropic cardiac and lung images, in comparison to the time the patient can hold his breath. Super-resolution methodology is applied to the multiple 2D slices acquired to improve the resolution of the images [7].

3.4.3 Content-based image retrieval

Content-based image retrieval (CBIR) is a technique for knowledge discovery, e.g., given a query image, the algorithm finds the most similar image in a given database. CBIR offers the possibility to identify similar case histories, understand rare disorders, and improve patient care. Extracting effective feature representations from pixel-level information and associating them with meaningful concepts are the main challenges in the development of CBIR methods. The CBIR community have gained interest in deep CNN models, due to their ability to learn rich features at multiple levels of abstraction.

To extract feature descriptors from medical images, all current approaches are utilizing (pre-trained) CNNs. For example, in [69] the authors trained a deep CNN to distinguish between different organs. From the evaluation dataset the features from the three fully connected layers in the network were extracted for the images. To retrieve the image, the same features were then extracted from the query image and compared with those of the evaluation dataset. According to [35] it is only a matter of time before CBIR can deliver successful application of deep learning methods. The direct training of deep networks for the retrieval task could be an interesting field of research.

3.4.4 Objective image quality assessment

The aim of objective image quality assessment of medical images is to classify image quality either as satisfactory or unsatisfactory for the following task. It is crucial to measure the objective quality of medical images to improve diagnosis and aid in better treatment [70]. In a recent study [71], image quality of fetal ultrasound was predicted using CNN. An attempt to reduce the data acquisition variability in echocardiograms using a CNN, trained on the quality scores assigned by an expert radiologist, was made in another study [72].

4 Methodology

The aim of this study is to investigate how medical imaging AI could be used in clinical practice. The ethical concerns of using AI in decision-making and diagnosis in medical practice will be studied. First, the choice of qualitative research will be discussed. Subsequently, the process of gathering data will be explained, and lastly, how the data was analyzed.

4.1 Research methodology

According to [73] qualitative research can be defined as "*any kind of research that produces findings not arrived at by means of statistical procedures or other means of quantification*". Instead of focusing on number, the focus lies on in-depth analyzing of experiences, opinions, and words. The individual is in focus when qualitative methods are utilized [74]. Qualitative research is generally associated with inductive reasoning, but a deductive approach can also be conducted.

The validity and reliability of qualitative methods are often questioned, these problems are more thoroughly discussed in point 4.1.1. The personal involvement of the researcher in an open study often contributes to perceiving qualitative research as rather subjective. Individual cases and rarely randomly picked samples are some of the reasons that contribute to limited generalization of qualitative research. However, the main goal of qualitative research is not generalizing to a population, but rather analyzing and grasping a specific case and context [75].

The choice of research methodology depends on the nature of the research question. This thesis is rather exploratory, and a qualitative research method appears to be suitable for this reason [6]. This thesis utilizes open questions to explore the interviewee's perspectives on AI in medical practice.

4.1.1 Reliability and validity of qualitative methods

The concepts of reliability and validity are advisable to consider, regardless of the choice of methodology. Reliability refers to whether a study is repeatable by a different researcher or by the same researcher at a different time [76]. Reliability is particularly difficult to achieve in qualitative studies. Reliability is often a challenge in qualitative interviews, due to the fact that the interaction between interviewer and interviewee is reflected under the circumstances

which the interview is conducted in [6]. Due to the changing context reproducing the interview might lead to a different outcome.

[76] introduces possible actions to cope with the problems of qualitative research, which have been adopted in this thesis. Choice of theory and research process must be displayed in a transparent way, allowing other researchers to follow the steps to grasp and reproduce the study. Furthermore, [76] stresses the importance of research report readers having access to the original data, not only generalizations and summaries. Considering these actions, transcribing the interviews and including direct quotes from the transcripts into the analysis are included in this thesis. Pre-testing the interview guide increases the reliability further [76].

Validity refers to whether a study accurately measured what it intended to measure [76]. In qualitative studies, particularly in research utilizing exploratory methods, this question is a bit more challenging to answer than in quantitative research. It is important that the observations fit with the theories for the validity of a qualitative study [75]. The quality of the design process of the study influences the validity of the research.

4.2 Gathering the data

To answer the research questions of this thesis, a suitable research method is required in order to gather data. Literature review and interviews appear to be a suitable method for this thesis. Literature review allows to collect data in a comprehensive way [77]**.** Interviews allow asking open-ended questions to the interviewees and probing experiences and opinions regarding the specific topic [76]. Interviews can be structured, semi-structured, unstructured, or an intermediate type. In unstructured and semi-structured interviews, the focus is on the interviewee's experience and opinion [6]. Thus, the interviews are rather informal, and interviewees are encouraged to speak freely about everything that comes to their mind. Exploratory studies are allowed to be less structured than confirmatory studies [76]. In a study like this thesis, with a small sample size, the focus is not on comparing the cases. Thus, the questions can be rather open and non-standardized.

Semi-structured interviews appeared to be a suitable method for this thesis. Structuring the interviews through an interview guide made it possible to keep orientation during the interviews. The structuring ensured that important theoretical topics were covered in the interviews. By conducting semi-structured interviews, individual experiences and opinions

were permitted to be shared in a non-constraining way. Due to the explorative nature of the research, the interviewees brought up topics that could not be foreseen. A structured interview might have missed out on these topics. The interview guide was highly flexible, and the questions could be asked in the most suitable order for the individual interview.

Interviews have limitations, as any other research method. Interviews are a social interaction and constrained by the specific interview situation [6].

4.2.1 Literature review

A systematic literature review was conducted in order to discover what is currently known, and what current models are based on regarding medical imaging AI and AI for healthcare. The publications were categorized as medical and non-medical (i.e. technical). Publications written in English were set as a criterion. Medical articles published after 2015 were included, as they were considered most relevant. An exception to the year criterion was made for [61], which revealed utilizing similar technology as presented in the newer publications. Non-medical articles published after 2013 were included in the review, to explain the techniques and technologies described in the medical articles. Focus for the medical articles was image recognition, but other solutions were also included for a broader perspective.

4.2.2 Interview guide

Before conducting the interviews, existing literature was studied intensively to gain insights about the technology and techniques behind AI for medical practice. The gained insights were used as an inspiration for open questions to understand the ethical concerns about using AI in medical practice. Relevant themes and categories were collected, and several questions were developed for each of the categories. Based on the questions, an interview guide was developed and interviews where customized according to the interviewee.

The questions were arranged into different categories in order to prepare fluent interviews. Due to the open and semi-structured nature of the interviews, the order of the questions in the interview guide was not strictly followed during the interviews. The interviewees were encouraged to answer the questions in an unconstrained way, allowing them to mention everything that came to their mind. Themes that were planned for a later part of the interview were brought up by some interviewees in the beginning, and naturally it seemed appropriate to advance to the questions regarding that theme. The interview guide served as a supporting

tool during the interview, to assure that all areas of the research were covered. The questions in the interview guide are presented in Table 2. The questions were chosen to probe and approach the research questions from different angles to establish a comprehensive understanding of the researched topic. The interviews put more emphasis on ethics, since the literature review revealed that the ethical concerns were absent in the articles.

Table 2 Interview questions

No.	AI
1.	How do you define AI?
2.	What is your experience of AI?
3.	What kind of AI solutions are you using in your daily work? • What kind of task is the AI doing? • Does the system require human intervention?
4.	Have you implemented AI at a client?
5.	What is the cost-benefit of using AI?
	Medical imaging
6.	What is your experience of AI in medical practice?
7.	How do you understand/interpret medical imaging AI?
8.	How can radiology benefit from AI?
9.	What kind of problems do you expect AI to solve in medical practice?
	Ethics
10.	Who is responsible for AI in general?
11.	Who is responsible when AI is used to support decision-making?
12.	Who is responsible for client data?
13.	How to ensure security and privacy of potentially sensitive data?
14.	How do you feel about using AI for diagnosis?
	Legal
15.	Is the legislation up to date?
16.	How to ensure security and privacy of potentially sensitive data?
	Doctors
17.	What kind of computer aid are you using now?

18.	How have those systems been helpful?
19.	Do you think patients are comfortable with having an automated diagnosis?
20.	In your opinion, which tasks can be handled by AI and which tasks should be done by a human? • Where do you draw the line?
21.	Are doctors ready to use AI in medical practice? • If yes, why do you think doctors are ready? • If no, what do you think it takes to get them ready?

4.2.3 Conducting the interviews

For this thesis, six persons were interviewed, three women and three men, as presented in Table 3. They were approached through different channels. Some were colleagues of the author, others were found through the supervisor's network. They were chosen through purposive sampling. In this sampling approach, the researcher tries to find cases rich in information providing certain attributes demanded by the character of the study [76]. For the present study, those attributes were related to the expertise of the persons, trying to have a sample covering a broad range of field, but also by their accessibility. All interviews have been conducted in English, one was audio-recorded, and the rest documented with notes during the interviews. Two interviews took place at the Accenture office in Heerlen. Four interviews were performed as Skype calls, because the researcher did not have the possibility to travel to Helsinki and Eindhoven for the interviews. Each interview took between 15 to 30 minutes.

Table 3 Personnel interviewed for these interviews

Case	Role	Description
A1	Data Engineering Manager	Working on a NLP project, has not implemented AI for a client.
A2	Application Development Associate Manager	Experience of image recognition in medical images, currently mostly working with data engineering and machine learning.
A3	Application Development Analyst	Data engineering experience, currently working on a NLP project for a client.
HUS1	Consultant Neurosurgeon	Working in clinical practice with neurosurgery, also involved in some research.
HUS2	A. Prof., Neurosurgery Resident	Working both in clinical practice and conducting research.

HUS3	Senior Consultant Anesthesiology and Intensive Care	Working mostly at the neuro intensive care unit with brain trauma patients. Also involved in research, current research focus on developing a tool to predict the outcome of trauma patients utilizing AI.

4.3 Analysis

The interviews were transcribed, after conducting them, to process them for the subsequent analysis. Qualitative content analysis appears the be a suitable method for the analysis. This method will be described next.

4.3.1 Qualitative content analysis

Qualitative content analysis is a systematic, rule guided qualitative text analysis [74]. Qualitative content analysis attempts to evolve from the strengths of qualitative analysis, such as utilizing guidance by rules and following the concepts of verification of reliability and validity. These strengths are then adjusted in a meaningful way by the method, for the analysis of qualitative data [74].

The data source needs to be defined as a first step to conduct qualitative content analysis. Presenting the interviewees, how the sample was chosen, what the conditions of the interviews were, and how the collected data was generated needs to be included in the first step, this is presented in section 4.2.3. The research questions that the study is based on, and the theoretical background must be completely clarified to define the intention of the analysis and to interpret the material [74]. Research questions and theoretical background were discussed in the theoretical part of the study (see section 1.1 and chapters 2 and 3) and were then included in the interview guide and the coding.

Summary, explication, and structuring are three basic forms of interpretation in qualitative content analysis [74]. Summary refers to the reduction of data, explication refers to finding additional data, and structuring refers to filtering out important features from the data. For this thesis, structuring and filtering out features appear to be most appropriate ways to analyze the data. For this analysis categories are defined, the categories in this thesis are AI, computer aid, and ethics. The interviews are filtered for statements fitting into the categories. The categories were developed inductively, led by the collected data. Some categories were determined in advance deductively, considering theoretical aspects in the development of the

questions. Most categories were revised during the analysis, as the topics of the interviews could not fully be predicted. To not influence the analysis with assumptions of the researcher this was an important procedure [74]. Disregarding data that is unfit for the categories is a risk when coding data [76]. Categories need to be defined carefully and checked for potentially important data falling outside the categories.

Two categories were split into subcategories and variables were developed for some of the subcategories, as seen in Figure 10. AI and ethics are split into several subcategories, and some of those subcategories have variables. To differentiate the categories from each other, coding rules were defined and supported by examples to guarantee a consistent analysis. All this merged into a coding agenda. The text was coded after establishing the categories and coding agenda.

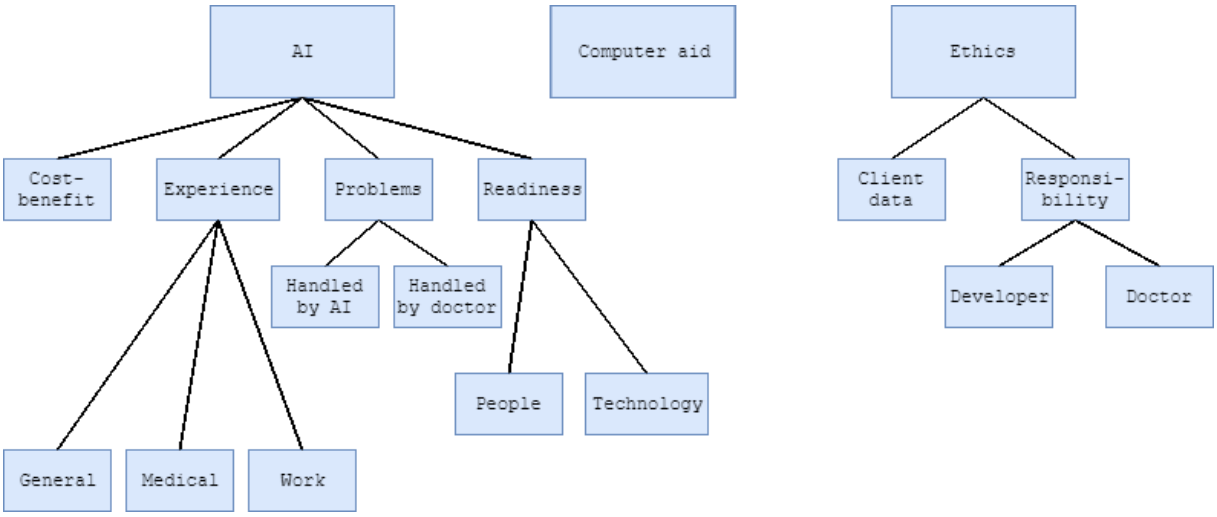

Figure 10 Categories.

4.3.2 Coding and content analysis

How the transcribed interviews were transferred into analyzable content will be described next. A coding agenda was developed first, where categories and variables were defined. Coding rules were developed to differentiate the categories. Next the transcripts were read, and relevant statements were underlined. A color scheme was utilized to structure the content of the transcript after reconsidering the defined categories. Statements, opinions and quotes were summarized into categories. New categories were developed when the contents did not fit in existing categories. Statements were summarized and rephrased by the researcher. The collected statements were analyzed and interpreted after coding all the

interviews, see chapter 5. Some categories were changed or merged to conduct the analysis, due to their correlation.

4.3.3 Discussion of the method

Qualitative content analysis considers both the context in which the data was generated and the theoretical background of the research [74], hence it appears to be a suitable method for this thesis. The steps of the analysis may be reconstructed due to the structured category system for coding the material. The reliability of the analysis and the comparability of the results are strengthened due to this. The category system could flexibly be revised during the process [74].

Studies based on very open research questions are not suitable for this method. Qualitative content analysis is not recommended for studies where an inductive development of categories might be restricting. The correlation between data and research questions guides the analysis, the rules and systems should not be inflexible [74]. Higher reliability is achievable when utilizing several researchers for coding. Due to the nature and time constraints of this thesis it was unrealizable to involve several researchers.

5 Literature review

An extensive literature review was conducted for this thesis. Six questions were posed for each article to prepare for the literature review. In total 38 articles were reviewed. The questions are presented in Table 4.

Table 4 Questions posed for the articles

No.	Questions
1.	What method is used in the solution?
2.	What is the background of the researchers?
3.	What is the state of the research? • Theory, or • Already in clinical use
4.	What is the purpose of the research?
5.	How is quality defined or assessed for the solution?
6.	Were any ethical concerns discussed in the article?

28 of the studies were conducted for a medical purpose, the remaining 10 were conducted for another purpose. 33 of the studies were conducted by universities and research institutions, and the remaining 5 were conducted by two private companies, Google and Microsoft. The background of the researchers was somewhat similar, mostly computer scientists, medical doctors, and engineers. Different fields of engineering were represented, but electrical and software engineers account for the largest share. The doctors were practicing at different departments with diverging specialties. 15 of the studies were conducted by teams consisting of both medical and technical (i.e. computer scientists, engineers, and other) researchers. 10 of the studies were conducted by only medical researchers and the remaining 13 studies were conducted by technical researchers. Of the 13 studies conducted by only technical researchers, 3 studies were conducted for medical purposes while the remaining 10 were conducted for other than medical purposes.

The methodologies used in 36 of the studies are similar, different forms and architectures of CNNs, many inspired by AlexNet [45], GoogLeNet [65], and VGG [46]. The two remaining studies are literature reviews that discuss CNNs. However, the architectures are diverse, and a common best performing architecture is not found among the studies.

None of the studies discuss the possible ethical concerns with implementing AI for clinical practice, the focus is only on the technical aspects. One of the studies briefly mentioned that implementing DL in radiology poses ethical challenges, and posed questions about the responsibility, but no solution to the problem was proposed. The ethical issues might be disregarded in the research as the studies reviewed are merely theories at this point. No evidence of practical usage of the studies was found. The studies conducted by Microsoft and Google might be in practical use, but no further information about their projects was found.

The studies offer solutions to a broad range of problems. The purposes of the studies were quite similar when considering the techniques, and diverse when considering the object and images that were utilized in the DL models. Both technical and medical studies focused on segmentation, classification, recognition, and detection of different images and objects. The objects of the images in the medical studies varied, e.g. lung nodules, brain tumors, and breast lesions, and different imaging modalities were utilized, e.g. MRI and CT scans.

5.1 Explanation of some abbreviations

The quality of the proposed solutions in the studies was assessed utilizing several different techniques. Determining which study has the best solution is difficult, since they all work with different datasets. However, a brief explanation of the quality assessment techniques is presented here. Area under the curve (AUC) is the area that is obtained by calculating a definite integral between two points. AUC is invariant to classification threshold and measures the quality of a model's prediction [78]. Intersection over union (IoU) is utilized to measure the accuracy of an object detector [79]. IoU is often used to evaluate the performance of SVM and CNN object detectors. IoU measures the overlap between the ground truth (i.e. the real object boundary) and the prediction. The Sørensen-Dice coefficient (SDC) compares the similarity between two samples. Mean average precision (mAP) is the average of the maximum precisions at different recall values [80]. Studies with no clear quality criterion are represented by –, in Table 5.

5.2 Summary of the literature review

Table 5 represents the literature, studied for the literature review. The first part of the table considers articles with focus on medical implementations of AI and the second part considers CNN research articles not particularly for medical implementations.

Table 5 Summary of the literature review

Articles	Method	Background of researchers	State	Purpose of research	Quality criteria	Ethics
Medical publications						
[72]	CNN-based regression model	Electrical and computer engineering, and medical	Theory	Reduce user variability in data acquisition by automatically computing a score of echo quality for operator feedback	Computation time, mean absolute error	No
[37]	Pretrained CNN	Medical	Theory	Develop a breast computer-aided diagnosis (CAD) methodology that addresses the issues by exploiting the efficiency of pre-trained CNNs and using pre-existing handcrafted CAD features	AUC	No
[52]	CNN	Medical	Theory	Evaluate the performance of a DL CNN model compared with a traditional NLP model in extracting pulmonary embolism findings from thoracic computed tomography reports from two institutions	Prediction accuracy (%), AUC	No
[41]	Pretrained GoogLeNet model	Computer science and medical	Theory	Present a CAD system for classifying thyroid nodules in ultrasound images	Classification accuracy (%)	No
[70]	Literature review	Electrical and electronic engineering	Theory	Review the recent advancement on Image Quality Assessment for medical images, mainly for MRI, CT and ultrasonic imaging	-	No
[59]	3d U-Net	Computer science and medical	Theory	Introduce a network for volumetric segmentation that learns from sparsely annotated volumetric images	IoU	No
[54]	CNN	Systems design engineering and medical	Theory	Present a segmentation algorithm for delineation of the prostate gland in DW-MRI via fully CNN	SDC	No
[58]	2d CNN, 3d CNN	Data science and medical	Theory	Present an automatic proximal femur segmentation method that is based on deep CNNs	AUC, SDC	No
[50]	CNN	Medical	Theory	Determine if deep learning, specifically CNN analysis, could detect and stage chronic obstructive pulmonary disease and predict acute respiratory	AUC	No

				disease events and mortality in smokers.		
[49]	CNN inspired by AlexNet	Computer information systems, electrical engineering and automation, and medical	Theory	Propose a novel approach that uses CNN for classifying brain medical images into healthy and unhealthy brain images	Classification accuracy (%)	No
[42]	Ensemble of CNN architectures	Biomedical and medical	Theory	Introduce a new method for classifying medical images that uses an ensemble of different CNN architectures	Classification accuracy (%)	No
[40]	Deep CNNs; AlexNet and GoogLeNet	Medical	Theory	Evaluate the efficacy of deep CNNs for detecting tuberculosis on chest radiographs	AUC	No
[47]	CNN	Electrical engineering and medical	Theory	Deep learning-based radiomics was developed to extract deep information from multiple modalities of MRIs	AUC	No
[35]	Literature review	Medical	Theory	Review the major deep learning concepts pertinent to medical image analysis and summarize over 300 contributions to the field	-	No
[61]	CNN	Medical	Theory	Developed several training methods in conjunction with a convolution neural network for general medical image pattern recognition	-	No
[7]	Literature review	Electrical and computer engineering, and medical	Theory	Review the clinical reality of radiology and discuss the opportunities for application of deep learning algorithms	-	Yes, briefly
[55]	Deep CNN	Technical and medical	Theory	Apply CNN to segmenting multiple sclerosis lesions and gliomas	SDC	No
[57]	CNN	Medical	Theory	An approach to 3D image segmentation based on a volumetric, fully CNN	SDC	No
[38]	Pretrained CNNs	Computer science and engineering, and medical	Theory	Apply a pretrained CNN to extract deep features from 40 CT images, with contrast, of non-small cell adenocarcinoma lung cancer, and combined deep features with traditional image features and	AUC	No

				trained classifiers to predict short- and long-term survivors		
[69]	Deep CNN	Computer and software engineering	Theory	Propose a framework of deep learning for content-based medical image retrieval system by using deep CNN that is trained for classification of medical images	Classification accuracy (%), mAP	No
[22]	Deep CNN	Information engineering and computer science	Theory	Propose some novel techniques, architectures, and algorithms to improve the robustness of distant talking acoustic models	-	No
[64]	Deep CNN	Medical	Theory	Develop a CAD system for masses in digital breast tomosynthesis volume using a deep CNN with transfer learning from mammograms	-	No
[60]	CNN	Electrical Engineering	Theory	Present three novel CNN-based architectures for glioma segmentation for images from the MICCAI BraTS Challenge dataset	SDC	No
[65]	CNN	Medical	Theory	Exploit three important factors of employing deep CNNs to computer-aided detection problems	-	No
[39]	Deep CNN	Engineering and medical	Theory	A DL with deep CNN and a non-deep learning with SIFT image feature and bag-of-word, a representative method for image recognition and analysis, were used to distinguish pathologically confirmed prostate cancer patients from prostate benign conditions patients with prostatitis or prostate benign hyperplasia	AUC	No
[71]	Two deep CNN models	Engineering and medical	Theory	To improve the efficiency of examination and alleviate the measurement error caused by improper ultrasound (US) scanning operation and slice selection, a computerized fetal US image quality assessment scheme is proposed to assist the implementation of US image	IoU	No

				quality control in the clinical obstetric examination		
[43]	CNN	Electrical and computer engineering, and medical	Theory	Determine whether deep learning models can distinguish between breast cancer molecular subtypes based on dynamic contrast-enhanced magnetic resonance imaging	AUC	No
[44]	CNN	Medical	Theory	Determine whether deep learning-based algorithms applied to breast MR images can aid in the prediction of occult invasive disease following the diagnosis of ductal carcinoma in situ by core needle biopsy	AUC	No
Non-medical publications, i.e. technical						
[51]	Literature review	Biomedical engineering; mathematics, informatics and mechanics	Theory	Provide insights into the design choices that can lead to efficient neural networks for practical application, and optimization of the often-limited resources in actual deployments	-	No
[62]	CNN	Google	Theory	A saliency-inspired neural network model for detection, which predicts a set of class-agnostic bounding boxes along with a single score for each box, corresponding to its likelihood of containing any object of interest	-	No
[67]	CNN inspired by VGG nets	Microsoft	Theory	Present a residual learning framework to ease the training of networks that are substantially deeper than those used previously.	Test and training error (%)	No
[53]	CNN	Engineering	Theory	Show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation	Pixel accuracy (%)	No
[36]	CNN	Computer science	Theory	Report on a series of experiments conducted for different recognition tasks using the publicly available code and model of the OverFeat network which	AUC and mAP	No

				was trained to perform object classification on ILSVRC13		
[68]	CNN	Microsoft	Theory	Introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals	Detection accuracy (%)	No
[46]	CNN	Engineering	Theory	Investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting	mAP	No
[66]	CNN	Google	Theory	Give clear empirical evidence that training with residual connections accelerates the training of Inception networks significantly	Top-5 error (%)	No
[63]	CNN	Google	Theory	Demonstrate that learning-based proposal methods can effectively match the performance of hand-engineered methods while allowing for very efficient runtime-quality trade-offs	mAP	No
[56]	CNN	Automation and engineering	Theory	Propose a new end-to-end network based on ResNet and U-Net	SDC	No

6 Results and analysis

The findings from the conducted interviews will be presented in this section. The findings will be analyzed and discussed in relation to the literature after each section of findings. The aim is, according to the research questions, to explore the interviewees' experience and opinions about utilizing AI in medical practice. The interviewees will be briefly presented, and the first part of discussion will be about their AI experience in general. Subsequently, the main aspects of possible implementations of AI in medical practice will be discussed. The chapter ends with a discussion about the ethical aspects of AI in medical practice.

6.1 Interviewee characteristics

The characteristics of the interviewees will be briefly presented to understand the backgrounds of the interviewed persons. Most of this information was collected during the interviews, some information was also collected through research of the interviewed persons.

In total, six persons were interviewed, three of them were female and three were male. Three of them are practicing physicians working for HUS in Helsinki, Finland and the other three are developers or managers working with IT (AI) for Accenture in Heerlen, the Netherlands.

6.2 AI

Since AI for medical practice is the main topic of this thesis, the topic is part of many of the categories in this analysis. At this point, the general experience of AI and the cost-benefit of possible implementations of AI in medical practice will be discussed.

6.2.1 Experience

To get an understanding of the interviewees experience of AI they were asked to briefly discuss what they know about the topic, if and how they are using AI personally and professionally. The interviewees from Accenture had naturally more experience of AI, since they are all working with it. Table 6 presents a summary of the interviewees' responses.

Table 6 AI experience

Case	Experience
A1	Using Siri (speech recognition) for help with simple tasks via iPhone at home. Currently working on an NLP project for work. No experience of AI in medical practice.

A2	Extensive experience of implementing AI in projects. Projects including extracting data from images with unstructured data with DNNs and CNNs, image recognition projects with CNNs. No experience of AI in medical practice.
A3	Extensive experience of implementing AI in projects for clients. More experienced in assembly learning (i.e. assembly language [81]), feature engineering, learning algorithms, and SVMs, than deep learning. Has not been involved in deep learning projects. Mostly focusing on developing NLP for client now.
HUS1	Not using AI in everyday practice. Has come in contact with AI through research projects.
HUS2	Not much experience of AI in medical practice since it is not commonly used yet in clinical practice at HUS.
HUS3	Not much experience of AI since it is not in clinical use yet at HUS. Currently involved in a research project where they are trying to predict the outcome of brain trauma patients utilizing a simple algorithm/AI.

Summing up the findings regarding the interviewees' experience, it was found that the Accenture interviewees possessed far more practical experience in AI than the doctors, which was also anticipated. The doctors conversely had limited experience of AI from research projects. All doctors were working in clinical practice and with limited focus on research. The limited experience of AI for the doctors might explain in part why AI is still not in use in clinical practice at HUS.

Only one of the interviewed Accenture employees had experience of AI for medical practice, from a project where the interviewee was working at another company. However, the other two had valuable insights from AI projects in other fields. The doctors were working at different units of HUS, which offered diverse insights. The Accenture employees were all from the same department, but due to working on different projects they could all offer important observations from different points of view.

6.2.2 Possible implementations

AI has great potential to benefit medical practice. Possible implementations and struggles were discussed with the interviewees and their responses are summarized in Table 7. Possible

implementations and prior research on deep learning in radiology were presented in chapter 3.

Table 7 Possible implementations and difficulties for AI in medical practice

Case	Possible implementations and difficulties
A1	(The topic was not discussed.)
A2	The lack of labeled data is a problem. It is important that the data gets labeled correctly.
A3	Hospitals have a lot of data, a lot of historical data. I think the problem is the lack of labeled data. You may have a huge set of CT scans, but you need someone to label those little points (irregularities on the scan) and sometimes you have some features that are not obvious to humans and then you need to have those to be able to train an AI to recognize those. It is a very repetitive task, basically you must go through thousands and thousands of cases and say this is A and this is B, which is not very efficient in this sense. Another problem is that you need partnership for medical imaging, with either hospitals or medical institutions for the data because they are the data sources in this case. I think in a lot of cases they are not allowed to share the data because of privacy issues. Patients should consent that their data is going to be shared for research purposes. I think most of the times patients are okay with that, the issues are due the law and the rulings. In the Netherlands, I am not sure, but I think that depends on the GDPR. It is a lengthy process.
HUS1	For example, it could be used to predict tumors in MRI and suggest treatment. There are good algorithms for image recognition. In Parkinson treatment we are inserting probes into the brain. AI could help to map precisely where to insert probe (precise target). But I think it is still a few years to come before AI will be in clinical and practical use in medical practice.
HUS2	For example, in image analysis you do not receive a radiological report immediately. The radiologist needs to put it together. With AI you could receive a report immediately and the quality of the report would not be dependent on the

	hospital or the radiologist. So, you could get a quality report at a small hospital, without a radiologist.
HUS3	A problem today in the ICU is that we have a lot of data, data is collected every second. But all that data is not being used since the human brain does not have the capacity to analyze all that data. There an AI would be extremely helpful.

AI has a great opportunity to benefit medical practice, extensive amounts of data are collected every day in medical practice. A large part of that data is left unused, since doctors do not have the time or capacity to analyze all of it. A problem with the data is that it is unlabeled, which contributes to overfitting and performance loss [17]. The AI developers may encounter difficulties to acquire medical data for training and testing of the models.

If the challenges can be conquered there are many possible implementations of AI for medical practice. Image analysis is a very repetitive and time-consuming task that could utilize an AI for predication and recognition. Tumor prediction was discussed in [7]. Automatic reports are another possible implementation, an AI could generate a report that was unbiased and independent of hospital and doctor. At any time of the day, a quality report could be received even if the specialist were not present or available.

6.2.3 Cost-benefit

Doctors spend a large amount of their day in front of a computer documenting, marking areas on scans, and completing mundane tasks [2]. By utilizing AI and other computer aid, mundane tasks could be handled by computers. Cost-benefit of utilizing AI were discussed with the interviewees and their responses are presented in Table 8.

Table 8 The cost-benefit of implementing AI for medical practice

Case	Cost-benefit
A1	If the process is highly complex and varying and requires human expertise it is not a good idea to bring in AI to that process. On the other hand, if the task/process is repetitive and mundane, implementing AI can be a good idea to speed up the process and free humans to handle more complex tasks.
A2	Benefits of AI, it really depends on which spectrum you are looking at. In predictive modeling it can help you with decision-making, if you have good quality data and

	you are able to make good predictions and you know in advance what should be your best bet in taking a certain decision. In medical imaging, for example, there are older studies in which AI implementations for retinopathy, cancer detection in CT scans and MRI images, where you have these trained models, with mostly deep learning, where of course this benefit is obvious. It supports doctors in detecting (diseases) in very early stages and detecting possible complications that could save life in the end.
A3	In processes with repetitive tasks AI can speed up the process since a robot is able to work 24 hours per day and weekends. That is not possible for humans.
HUS1	(The topic was not discussed.)
HUS2	AI could be very beneficial, for example the AI could do the initial screening so there would be no need to have specialist around at night (24/7). I also think AI can save time and make like a summary of all findings, so doctors have more time to interact and treat patients and not sit with documents in front of the computer.
HUS3	It is a difficult question. I am currently developing a simple model that can be installed on any computer worldwide, it doesn't require any extra or special computer power. That will hopefully save money. I don't think it will give the doctor more time for patient interaction. I think it will require more thinking to understand what the model is doing. In the ICU the traumas are so complex that it will not add any time for interaction.

The benefit should be greater than the cost of implementing AI, however, doctors need training to be able to use AI. Initially an investment is needed, but the return on investment might be significant. Computer aid is already in use at HUS, e.g. EHR has been utilized already for 15 years in Finland. All data is collected and stored electronically, an advantage for AI implementations.

One of the interviewees sees AI as a tool to save money and free up time for patient interaction. Another doctor does not believe in gaining more time for patients, instead he expects that the AI requires more thinking from the doctors to fully understand what the AI is doing. The benefit of AI is that it can work without breaks and that it is convenient when repetitive tasks are at hand.

6.2.4 Readiness

Implementation is one part of bringing AI to medical practice, the other part is to assure that the users of the AI are ready. The doctors interviewed for this study did not have extensive experience of AI, considering that they also were young I presume that the older doctors will not have any more experience. Due to the lack of experience AI might appear complicated and intimidating at first sight. Training for the doctors is an essential part of building trust and knowledge of AI. Table 9 presents the responses of the interviewees.

Table 9 Readiness of the people

Case	Readiness
A1	Using AI as an aid is fine, but a doctor should always oversee the AI and the diagnosis should be made by utilizing the knowledge and experience of the doctor.
A2	Personally, I think it is a big boost for medical practice and it could really result in benefits for a lot of patients. I am strongly supporting that, I think though an AI will not work solely so the AI is there to support doctors into spotting some cases that may not be so obvious for them. So, I think it is like any other tool that you are developing for making your life easier. You had scissors in the past and then you started developing surgical knives (=scalpels) to help them. There are a lot of tools developed to help people and AI is just another tool.
A3	I think AI is very helpful to use in diagnosis, but the user needs to be informed about the data the AI is operating on. The AI can work on its own, but it needs to be controlled and monitored by doctors/experienced personnel. The decision needs to be controlled by the doctor. Well-labeled data is a huge asset for implementing AI in medical practice. If all the hospitals and medical institutions would work together a very valuable combined knowledge would be achieved, imagine some sort of medical internet database and how the medical practice could benefit from that (compare with ImageNet).
HUS1	Yes, or at least the young ones (are ready to use AI in clinical practice). The older generation might be frightened by AI and will want to do things as they always have done. But there is also the risk that doctors will start to put too much trust in AI and blindly follow the prediction they are given.

HUS2	Doctors are not yet ready (to use AI in clinical practice), I think it still needs academic and scientific background (research) to make doctors trust AI. Before bringing anything new (models) into medical practice extensive trial is done. I do not think AI can be tested in the same way as we always tested new models but it needs more research and academic backup before it is ready for medical practice. I think doctors are still skeptical about using AI. I am very positive, and I am looking forward to using AI. But as I said, it still needs a lot of research to back it up.
HUS3	I don't think you need to say it to the patient that a computer is making the diagnosis. You should inform them that it is just another tool to help the doctors and, in the end, it is the doctor that makes the decision for the diagnosis. It is also important to inform the doctors about the ability and limitations of the AI.

Younger doctors might be ready for AI tools, one interviewee concluded. Another interviewee claimed that doctors are not ready, pointing to the fact that there is too little research in the field for doctors to trust the technology. Training is necessary to prepare doctors for AI and informing the doctors about the capability and limitations of the system. A critical attitude against the AI might be healthy, trusting the AI blindly and not questioning the predictions might have fatal consequences. The patients should be informed about the AI accordingly.

All interviewees concluded that AI has extensive potential for medical practice. The possible implementations were discussed thoroughly in chapter 3. More research, proof of concept and clinical trials are essential to build trust for AI among doctors. Trust can be added by including doctors in the development process. The interviewed doctors all appeared hopeful and eager to include AI tools in clinical practice.

6.3 Ethics

AI in medical practice is surrounded by promises and excitement. It has the ability to improve the quality of medical practice, reduce errors in predictions and diagnosis, make processes more efficient, and help prevent human error and bias. There are many promises, but who is

responsible for the decisions made by AI? That is one of the ethical questions that need to be considered for the benefits of AI to be realized.

AI algorithms can be implemented to cheat and perform in unethical ways, as proven in Uber's Greyball algorithm [82] for deceiving law enforcement officials in cities where Uber's services violated regulations. Volkswagen used an algorithm to enable vehicles to pass emission tests by reducing nitrogen oxide emissions during tests [83]. The values of the AI developers are not always aligned with the values of the doctors [84]. Systems can be guided to improve quality metrics while ignoring patient care. The algorithms might be able to alter the provided data when regulators are reviewing the hospital. Another temptation is to develop clinical decision-support systems to generate increased profits for the developers or users, by recommending tests, drugs, or devices in which they are stakeholders.

6.3.1 Responsibility

Only the responsibility of when AI is used in decision-making was discussed with the interviewees, the responsibility for ethical development of the system was not discussed. Table 10 shows the interviewees' opinion on who is responsible when AI is used in decision-making.

Table 10 Responsibility when AI is used for decision-making

Case	Responsibility
A1	The one that uses the AI to execute processes is responsible (in this case that is the doctor). Accenture has a framework called Responsible AI that defuses the ethical risks with AI, humans are the ones that control the AI, making sure that biased data is not part of the decision process/predictions.
A2	It is the responsibility of the doctor for example in this case, to make sure that (s)he does not use the system in a wrong way with maybe not good enough data and then the doctor blindly trusts the algorithm and does not doublecheck (the diagnosis). The developers of course are responsible for documenting exactly the type of process the AI is doing. As an AI developer you are supposed to let the people that

	are going to work (with the system that you are training) to know exactly what the boundaries of the model are, what the model can do and cannot do. The developer is responsible to make it very clear that it is not going to provide the user the answer to everything, to document on what type of data (the model needs), the minimum requirements of the data and what are the core prerequisites of making that AI to work.
A3	Everyone who is involved in the process of developing the AI (are responsible), the developers have the greatest responsibility, but it should be everyone's concern. The user must be informed how to use the AI and the limitations of it.
HUS1	Doctors should be responsible, but it might be the company that developed the AI that can decide too much. E.g. treatment and protocols differ from country to country, so doctors must take the most responsibility.
HUS2	The doctors are responsible for AI and they must also be able to trust it.
HUS3	I think that doctors should be involved in implementing AI for medical practice, if an external party is developing the AI the doctors do not have full understanding of how it works and the limitations of the AI.

Both developers and doctors are responsible for the system, but the responsibilities differ for the parties. The doctors for their part are responsible for using the system right and providing good quality data to the system. The developers are responsible for informing how the system should be used, what type of data it needs, the minimum requirements, and the limitations of it. The substantial responsibility lies on the doctors when the system is delivered, while the developers hold the largest responsibility during the development process to deliver an ethical system. An ethical system in this context refers to a system that does not act on biased data or discriminate on a certain basis, e.g. the algorithms in [82, 83] were unethical.

The doctors need to be involved in the development process, to gain insight about the system's capability, and to guarantee that the system is developed according to the doctors' needs. The responsibility appears to be divided between developers and doctors during the whole process.

One interviewee explained how Accenture is coping with developing ethical AI, via their framework called 'Responsible AI'. It is essential that the ethical concerns are considered when AI is developed. Biased data should not be used for predictions or decision-making; if biased data is used it leads to a biased outcome.

6.3.2 AI vs doctor

AI promises to handle a wide range of tasks. The question which tasks can be handled by AI and which tasks need to be handled by a doctor were discussed with the interviewed doctors. Table 11 presents a summary of the doctors' opinions on this topic.

Table 11 Tasks handled by AI or doctors

Case	AI vs doctor
HUS1	Calculations can be handled by AI. For example, to find correlations between 1000 MRIs can be done by AI. AI can be used to help with predictions, prognosis, and diagnostics. Patient care and interactions to see how the patient is feeling must be done by doctors. Statistics do not tell the whole truth. Treatment decisions should also be taken by doctors.
HUS2	For example, image analysis: you do not get a radiological rapport immediately. The radiologist needs to put it together. With AI you could get a rapport immediately and the quality of the rapport would not be dependent on the hospital or the radiologist. So, you could get a quality rapport at a small hospital, without a radiologist. AI could do the initial screening so there would be no need to have specialist around at night (24/7). Documentation (objective) can be done by AI and screening methods and suggestions could be handled by AI. Decision can be supported by AI. The radiologist confirms the prediction. Doctors should interact and treat patients. The final decision should be taken by the doctors.
HUS3	All simpler tasks in patient treatment can benefit from an AI.

	The final decision must be made by a doctor. In the ICU there are a lot of patients with complex traumas, there I think a human is more suitable to handle the treatment and plans.

The interviewees suggested different tasks that the AI, in their opinion, could handle, e.g. simple calculations, reporting, documenting, and support in decision-making. Their opinions were somewhat differing, but they concluded unanimously that the final decision should be taken by the treating doctor.

6.3.3 Client data

The client data, or in this case the patient data, is sensitive and privacy needs to be secured. The ownership of client data and how to ensure that it is secure and not used for malicious purposes were discussed with the interviewees, see Table 12.

Table 12 Responsibility for client data

Case	Client data
A1	The owner of the data should be the one collecting and using the data, i.e. the hospital. The AI provider should never own the data. Data protection is ensured due to the GDPR legislation.
A2	I signed a waiver for like giving my data, and there are some clauses in the consent that I signed, right. I am just assuming that those clauses are respected and if those clauses are not respected that will dig me in to trouble at some point, like my data being leaked out. Once there is a uniform consent that both parties have signed, both parties should then commit to that contract. For me it is pretty straight-forward in that sense, but of course you never have the certainty like once it goes there like you don't have a track of who is looking at your data
A3	GDPR and regulations take care of the security of client data nowadays. At least at a big company such as Accenture, complying with the GDPR is done automatically. Also, when a contract with a client ends, all the client data is deleted.
HUS1	The owner of the data should be the institution or the hospital and the patient it concerns. AI provider should not have the right to the client/patient data.
HUS2	The hospital or institution that is using the AI is responsible and the owner of the data. It is very unethical to sell the patient data to the AI developer, at least in Finland.

HUS3	The treating doctor has the main responsibility for patient data. We follow a strict protocol for handling patient data, it is only the treating doctor (and nurses) that have access to the data. You also need approval from the ethical committee before taking out any data for a research or another project. So, for our part I think that the data is secure and cannot be misused.

The interviewees concluded that client data is secured in the European Union by the General Data Protection Regulation (GDPR) and is protected from misusage. The client data should belong to the one that is collecting the data, in this case the institution or the hospital. The developer of the system should not have ownership of the data. Hospitals follow strict protocols for handling client data and only a few have access to it.

6.4 Discussion

As discussed above, AI promises extensive opportunities to improve and streamline clinical practice. Doctors spend more time in front of computers doing mundane and repetitive tasks than interacting with patients [2]. AI can handle repetitive tasks and free up time for doctors to do more complex tasks.

Extensive amount of data is collected and documented every day in clinical practice, but due to the limited capacity of the human brain and time the doctors cannot analyze all of it. AI has the capacity to analyze large datasets. Image analysis is suitable for AI, e.g. determining if a tumor is malignant or benign, as presented in [7, 37]. The findings of this study established that medical imaging is suitable for AI, for time-consuming tasks like object and lesion detection [35]. Another possible task was found to be reporting, unbiased and independent of hospital and doctor.

The doctors need to trust and be ready for AI, which the findings showed that they are not. Training can prepare the doctors for AI. It was discovered that more research is necessary for building trust amongst doctors. Well-informed doctors and critical thinking appears to be essential for implementing AI in clinical practice. Findings concluded that the interviewed doctors are aware of the promises and limitations of AI.

The findings also revealed some ethical concerns about bringing AI into clinical practice. Responsibilities need to be determined in case a faulty decision is made. A concern is the implementation of unethical AI and learning biases, wrongful usage of AI was discussed in [83,

82]. Ignoring patient care in favor for improved quality metrics should be avoided. The most decision should be made by a doctor, and AI can be used for support in decision making and predictions. Blindly trusting the AI is not an option, more critical tasks should be handled by doctors, according to the findings. The developers for their part are responsible for developing an ethical and unbiased system. Including doctors in the development process is found to build more trust for the AI.

Security for the client data is ensured with GDPR in the European Union and all companies are expected to comply with that regulation.

6.5 Limitations and reliability

Due to the breadth of the field of AI and medical imaging AI possibilities, as well as the exploratory character of this thesis, it intended to, and could, only contribute with a small insight. To truly understand the challenges and ethics of implementing AI in medical practice, the area would need to be studied from different angles and more in depth. A larger sample and a more diverse one could provide a more thorough explanation of the problem. For this reason, the limitations and reliability of this thesis are discussed.

The study is limited by the number of persons interviewed. The fact that employees from only two companies, Accenture and HUS, were interviewed might also have an impact on the result. For future research, comparing the results of this study with findings from another similar study appears recommendable. The same research with a bigger and more diverse sample might present insights into the topic and underlying processes of implementing AI in medical practice, as well as the results regarding readiness of the technology and doctors.

According to [75], the researcher is the most important tool in a qualitative survey, which makes the research dependent on where the researcher places her/his emphasis. In other words, one researcher may put emphasis on one topic while another researcher believes that another topic is more important. In many cases, it is thus considered impossible to replicate qualitative studies [75].

7 Conclusion

This thesis studied the readiness of including AI in medical practice and the opportunities and challenges it is facing. Based on the literature and findings analyzed above, the main results of this thesis are summarized and recommendations for possible future research are discussed. Conclusion will be made regarding two main points, the readiness of the technology and the doctors, and the ethical aspects that need consideration.

The main results of this thesis discovered that different architectures of CNNs are most commonly used in the reviewed studies. The proposed CNNs all perform reasonably well during the studies and on the tested datasets. The studies focus on the technical aspects and the discussion about ethical dilemmas are missing.

AI is not in clinical use yet at HUS. The interviewed doctors presumed that it is still a few years away before AI is used in day to day practice at HUS. The doctors lack experience of AI and need trainings to build trust and understanding for the system. More research and clinical trials is a must to gain readiness amongst doctors. The doctors are responsible for handling the AI correctly and feeding it good data. The developers are responsible for documenting and developing an unbiased system. Ethical development and usage need to be assured, how is still a question. Radiology would be a good starting point for including AI in clinical practice. Medical imaging could greatly benefit from DL tasks, such as classification, segmentation, and detection and could be monitored by doctors, to assure that correct decision and predictions are made.

Due to the limited time of this thesis, it could only contribute with a small insight. Future research could further investigate the ethical concerns when implementing AI for clinical practice and possible conduct a case study at a hospital or medical institution. Another possible future research could study and develop an ethical framework for AI in medical practice, involving both doctors and developers.

8 Sammanfattning

Introduktion

En mängd uppgifter utförs i dagens samhälle med hjälp av artificiell intelligens (AI), till exempel väderleksrapporten och ansiktsigenkänning tack vare framsteg inom datavetenskapen och ultrasnabba beräkningshastigheter. Vilken inverkan AI kommer att ha på sjukvården är ännu okänt. Maskiner lär sig att upptäcka mönster som inte kan avkodas med hjälp av biostatistik genom att processera stora dataset (eng. *big data*) genom matematiska modeller (algoritmer). AI har framgångsrikt testats för bildanalys i radiologi, patologi och dermatologi. Det har resulterat i snabbare diagnoser med samma noggrannhet som medicinska experter producerar. Det är svårt att nå en 100% diagnostisk tillförlitlighet, men genom att kombinera maskiner och läkare kan systemets prestanda förbättras. Kognitiva program påverkar medicinsk praxis genom att tillämpa NLP för att läsa den snabbt växande vetenskapliga litteraturen och granska åratal av elektroniska medicinska journaler. Genom att implementera AI i sjukvården kan medicinska fel minskas och vården för patienter med kroniska sjukdomar kan förbättras [1].

Forskningsfrågorna i denna avhandling är:

1. Används AI i medicinsk bildvetenskap?
2. Vad kan förväntas i framtiden?
3. Är tekniken utvecklad tillräckligt för att vara pålitlig och kunna användas i sjukvården (utan vidare utveckling)?

Enligt forskningsfrågorna är huvudsyftet för denna avhandling att undersöka hur AI för medicinsk bildbehandling idag används i klinisk praxis. Två stödjande forskningsfrågor definierades för att ytterligare avgränsa forskningsfokuset. De två stödjande frågorna avser att undersöka vilken inverkan klassificeringsteknik kommer att ha i sjukvården och om tekniken redan nu är redo att används utan vidare utveckling.

Utgående från en litteraturstudie och intervjuer syftar forskningen till att utvärdera möjligheten att använda AI i medicinsk praxis samt undersöka de etiska perspektiven när AI används i sjukvård. Möjliga implementeringar diskuterades även med intervjuobjekt för att få en bredare förståelse om ämnet. Totalt studerades 38 artiklar och några frågor ställdes till

varje artikel. Med frågorna försöker studien avslöja hur AI och djupinlärning (*deep learning*, DL) används i sjukvården.

Medicinsk bildvetenskap

Medicinsk bildvetenskap (eng. *medical imaging*) består av tekniker och processer för att visualisera kroppsdelar, vävnader eller organ för medicinska ändamål, t.ex. diagnostisera sjukdom och planera behandling. Medicinsk bildvetenskap inbegriper discipliner som radiologi, endoskopi, mikroskopi, bildbehandling och visualisering. För att begränsa omfattningen av denna avhandling kommer den att fokusera på radiologi, eftersom det finns många artiklar och omfattande litteratur inom området. Radiologi innebär avbildning av människokroppens inre organ för kunna fastställa diagnoser.

Djupinlärning och radiologi

Ett av de första områdena där DL gjorde en signifikant förbättring för medicinsk analys var bildklassificering. DL är för tillfället den mest använda metoden för klassificering av radiologiska data. CNN-nätverk med varierande arkitektur används av de flesta tillgängliga DL-metoder. Jämfört med naturliga bilder, som har drivit utvecklingen av DL-tekniken under de senaste fem åren, är tillgängligheten av radiologiska data begränsad. DL behöver ett stort antal bilder för att tränas.

Kvantitativ analys av kliniska parametrar relaterade till volym och form, till exempel hjärt- eller hjärnanalys, aktiveras genom segmentering av organ och andra understrukturer i medicinska bilder. Segmentering är ett viktigt första steg i datorstödd detektering [35]. I både naturlig och medicinsk bildanalys är segmentering en frekvent uppgift. En bild delas upp i olika regioner för att klassificera varje pixel i bilden med ett CNN-nätverk. De vanliga applikationerna är segmentering av organ, understrukturer eller lesioner i radiologi, i allmänhet som ett förbehandlingssteg för extraktion och klassificering [7]. Ett viktigt förbehandlingssteg vid segmentering eller i det kliniska arbetsflödet för behandlingsplanering är lokalisering av anatomiska objekt, såsom organ eller landmärken (eng. *landmarks*). Parsning av 3D-volymer krävs ofta för lokalisering inom medicinsk bildvetenskap. Flera metoder har föreslagits för att underlätta 3D-dataparsning genom att implementera DL, t.ex. att behandla 3D-rymden som komposition av 2D ortogonala plan. Den mest populära strategin som har uppvisat goda resultat för att identifiera organ, regioner och landmärken är lokalisering genom 2D-

bildklassificering med CNN-nätverk. Enligt [35] utförs forskning på detta koncept med betoning på exakt lokalisering genom att modifiera inlärningsprocessen. Författarna förutspår att sådana strategier kommer att undersökas ytterligare i framtiden.

Resultat

En omfattande litteraturstudie genomfördes för denna avhandling. Sex frågor ställdes till varje artikel. Totalt granskades 38 artiklar. Frågorna presenteras i tabell Table 4.

Av de 38 granskade artiklarna låg fokus på att lösa något medicinskt problem i 28 av dem, medan de resterande 10 artiklarna fokuserade på att lösa ett mer generellt problem. 33 av studierna genomfördes av universitet och forskningsinstitut, och de övriga 5 genomfördes av två privata företag, Google och Microsoft. Forskarnas bakgrund var rätt likartad, mestadels datavetare, läkare och ingenjörer. Olika teknikområden var representerade, men el- och mjukvaruingenjörer står för den största andelen. Läkarnas bakgrund var rätt varierande, flera olika områden och specialiteter var representerade. 15 av studierna genomfördes av team bestående av både medicinska och tekniska (dvs. datavetenskapare, ingenjörer och andra) forskare. 10 av studierna utfördes av endast medicinska forskare och de övriga 13 studierna utfördes av tekniska forskare. Av de 13 studier som utförts av enbart tekniska forskare genomfördes 3 studier för medicinska ändamål medan de återstående 10 utfördes för andra än medicinska ändamål.

Metoderna som används i 36 av studierna är liknande, olika former och arkitekturer av CNN-nätverk (eng. *convolutional neural network*), många inspirerade av AlexNet [45], GoogLeNet [65] och VGG [46]. De två återstående studierna är litteraturrecensioner som diskuterar CNNs. Arkitekturerna är emellertid olika, och en allmän arkitektur finns inte bland studierna.

Ingen av de studerade artiklarna diskuterar möjliga etiska problem med AI i klinisk praxis, fokus ligger endast på de tekniska aspekterna. En av studierna nämnde kortfattat att implementering av DL i radiologi är etiskt svårt och vem som har ansvaret frågades, men ingen lösning på problemet föreslogs. De etiska frågorna ignoreras kanske i forskningen, eftersom de granskade studierna än så länge bara är teorier. Inga bevis på praktisk användning av studierna hittades. Undersökningarna som genomförts av Microsoft och Google kan vara i användning i praktiken redan, men inga ytterligare uppgifter om deras projekt hittades.

Studierna erbjuder lösningar på ett brett spektrum av problem. Både de tekniska och de medicinska studierna fokuserade på segmentering, klassificering, erkännande (eng. *prediction*) och detektering av olika bilder och objekt. Objekten i de medicinska studierna var varierande, t.ex. lungnoduler, hjärntumörer och bröstskador, och olika avbildningsmodaliteter användes, t.ex. MR- och CT-skanningar.

Intervjuerna avslöjade att intervjuobjekten från Accenture hade mycket mer praktisk erfarenhet av AI än läkare**,** vilket också var förväntat. Läkarna hade å andra sidan kommit i kontakt med AI i olika forskningsprojekt. Läkarnas begränsade erfarenhet av AI kan delvis förklara varför AI fortfarande inte används i klinisk praxis vid HUS. Endast en av de intervjuade Accenture-medarbetarna hade erfarenhet av AI för medicinsk praxis, från ett projekt där intervjuobjektet arbetade för ett annat företag. De två andra hade dock värdefulla insikter från AI-projekt inom andra områden. Läkarna arbetade på olika enheter av HUS, vilket gav olika insikter. Intervjuobjekten från Accenture arbetade på samma avdelning, men på grund av att de arbetade med olika projekt kunde de alla erbjuda viktiga observationer ur olika synvinklar.

I teorin borde fördelarna vara större än kostnaden för att implementera AI i sjukvården, men läkarna behöver skolning för att kunna använda AI. Initialt krävs en investering, men avkastningen på investeringen kan vara betydande. Datorer används i bred utsträckning redan vid HUS, t.ex. elektroniska patientjournaler har redan använts i Finland i 15 år. Alla data samlas in och lagras elektroniskt, vilket är en fördel för framtida AI-implementeringar.

Yngre läkare kan vara redo för AI, konstaterade ett intervjuobjekt. En annan hävdade att läkare inte är redo eftersom det inte har forskats tillräckligt inom området för att läkare ska lita på tekniken. Skolning är nödvändig för att förbereda läkare för AI och informera läkarna om systemets förmåga och begränsningar. En kritisk inställning mot AI kan vara hälsosam, att blint lita på AI och inte ifrågasätta prognoserna (eng. *prediction*) kan leda till fatala konsekvenser. Alla intervjuade konstaterade att AI har stor potential i klinisk praxis. Mer forskning och kliniska prövningar är avgörande för att bygga förtroende för AI bland läkare.

Både utvecklare och läkare är ansvariga för systemet, men ansvaret skiljer sig åt för parterna. Läkarna är ansvariga för att systemet används rätt och att systemet förses med kvalitativa data. Utvecklarna ansvarar för att informera användarna om hur systemet ska användas,

vilken typ av data den behöver, minimikraven och begränsningarna för systemet. Det yttersta ansvaret ligger hos läkare när systemet levereras, medan utvecklarna har det största ansvaret under utvecklingsprocessen för att leverera ett etiskt system. Ett etiskt system i detta sammanhang avser ett system som inte agerar på partiska (eng. *biased*) data eller diskriminerar på en viss grund, t.ex. i [82, 83] var algoritmerna oetiska. Läkarna måste vara involverade i utvecklingsprocessen, för att få insikt om systemets förmåga och för att garantera att systemet utvecklas enligt läkarens behov.

I intervjuerna framgick att patientdata skyddas av dataskyddsförordningen (GDPR) och borde därför inte kunna missbrukas. Patientdata ska tillhöra den som samlar in data, i det här fallet institutionen eller sjukhuset. Sjukhus följer strikta protokoll för hantering av patientdata.

Slutsats

Denna avhandling studerade vilka möjligheter och utmaningar AI ställs inför i sjukvården. Baserat på den studerade litteraturen och fynden som analyseras ovan, sammanfattas de viktigaste resultaten i denna avhandling och rekommendationer om möjlig framtida forskning diskuteras. Slutsatsen kommer att fokusera på tekniken, läkarnas beredskap och de etiska aspekter som behöver beaktas.

Olika arkitekturer av CNN-nätverk är främst förekommande i den studerade litteraturen. De föreslagna CNN-nätverken presterar relativt väl i testerna och på testdata. Studierna fokuserar endast på de tekniska aspekterna och diskussion om etiska dilemman saknas.

AI är inte i klinisk användning vid HUS. Läkarna som deltog i intervjuerna antog att det fortfarande är några år kvar innan AI kommer att börja användas dagligen vid HUS. Läkare saknar överlag erfarenhet av AI och behöver utbildning för att få förtroende och förståelse för systemet. Mer forskning och kliniska studier behövs för att göra läkarna redo för AI. Läkarna är ansvariga för att AI hanteras korrekt. Utvecklarna ansvarar för att dokumentera och utveckla ett opartiskt system. Hur ett etisk AI ska utvecklas och användas är ännu ett frågetecken. Radiologi skulle vara en bra utgångspunkt för att inkludera AI i klinisk praxis. Medicinsk bildbehandling kan i stor utsträckning dra nytta av DL, såsom klassificering, segmentering och detektion och kan övervakas av läkare, för att säkerställa att korrekta beslut fattas.

Denna avhandling kunde endast bidra med en liten inblick i ämnet, på grund av den begränsade tidsramen. Framtida forskning kunde ytterligare undersöka de etiska problemen vid implementeringen av AI i klinisk praxis och eventuellt genomföra en fallstudie på ett sjukhus eller en medicinsk institution. En annan möjlig framtida forskningsstudie kunde studera och utveckla en etisk ram för AI inom medicinsk praxis, genom att involvera både läkare och utvecklare.

References

[1] E. W. Brown and D. D. Miller, "Artificial Intelligence in Medical Practice: The Question to the Answer?", *The American Journal of Medicine,* pp. 129-133, 2018.

[2] S. Vihavainen, "Kysely: Kolmannes lääkäreistä käyttää yli kuusi tuntia työvuorosta tietokoneisiin – ”Joka hetki 600 lääkäriä tuijottaa ruudulla olevaa tiimalasia”“, *Helsingin Sanomat,* 2016.

[3] I. Sample, "'It's going to create a revolution': how AI is transforming the NHS”, 4 July 2018. [Online]. Available: https://www.theguardian.com/technology/2018/jul/04/its-going-create-revolution-how-ai-transforming-nhs. [Accessed 17 July 2018].

[4] Microsoft, "Project InnerEye – Medical Imaging AI to Empower Clinicians”, 2008. [Online]. Available: https://www.microsoft.com/en-us/research/project/medical-image-analysis/. [Accessed 19 July 2018].

[5] K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy, E. W. Karlson, A. N. Ananthakrishnan, V. S. Gainer, S. Y. Shaw, P. Szolovits, S. Churchill and I. Kohane, "Development of phenotype algorithms using electronic medical records and incorporating natural language processing”, *BMJ,* 2015.

[6] M. Saunders, P. Lewis and A. Thornhill, *Research methods for business students*, Harlow: Pearson Education Limited, 2009.

[7] M. A. Mazurowski, M. Buda, A. Saha and M. R. Bashir, "Deep learning in radiology: an overview of the concepts”, 10 February 2018. [Online]. Available: https://arxiv.org/abs/1802.08717. [Accessed 25 July 2018].

[8] A. Bernasconi, "Structural Analysis Applied to Epilepsy”, in *Magnetic Resonance in Epilepsy*, 2nd ed., Academic Press, 2005, pp. 249-269.

[9] "Dogs vs. Cats”, [Online]. Available: https://www.kaggle.com/c/dogs-vs-cats. [Accessed 26 June 2018].

[10] B. Romera-Paredes and P. H. S. Torr, "Recurrent Instance Segmentation”, 25 November 2015. [Online]. Available: https://arxiv.org/abs/1511.08250. [Accessed 20 July 2018].

[11] A. Ouaknine, "Review of Deep Learning Algorithms for Object Detection”, 5 February 2018. [Online]. Available: https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852. [Accessed 28 June 2018].

[12] C. Woodford, "Neural networks”, 14 March 2018. [Online]. Available: https://www.explainthatstuff.com/introduction-to-neural-networks.html. [Accessed 31 August 2018].

[13] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara and A. Hampl, "Artificial neural networks in medical diagnosis", *Journal of Applied Biomedicine,* pp. 47-58, 2013.

[14] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press Ltd, 2016.

[15] U. Karn, "An Intuitive Explanation of Convolutional Neural Networks", 11 August 2016. [Online]. Available: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/. [Accessed 26 June 2018].

[16] A. Karpathy, "CS231n: Convolutional Neural Networks for Visual Recognition", [Online]. Available: http://cs231n.github.io/neural-networks-1/#nn. [Accessed 31 August 2018].

[17] D. Bacciu, P. J. G. Lisboa, J. D. Martín, R. Stoean and A. Vellido, "Bioinformatics and Medicine in the Era of Deep Learning", 27 February 2018. [Online]. Available: https://arxiv.org/pdf/1802.09791v1.pdf. [Accessed 19 June 2018].

[18] G. Wu, M. Kim, Q. Wang, B. C. Munsell and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning", *IEEE Transactions on Biomedical Engineering,* pp. 1505-1516, 2016.

[19] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen and C.-M. Chen, "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary noduels in CT scans", *Scientific Reports,* 2016.

[20] S. Lau, "Image Augmentation for Deep Learning", 10 July 2017. [Online]. Available: https://towardsdatascience.com/image-augmentation-for-deep-learning-histogram-equalization-a71387f609b2. [Accessed 14 November 2018].

[21] "About ImageNet", 2016. [Online]. Available: http://image-net.org/about-overview. [Accessed 14 November 2018].

[22] M. Ravanelli, "Deep Learning for Distant Speech Recognition", 2017. [Online]. Available: https://arxiv.org/abs/1712.06086.

[23] K. B. Ahmed, L. O. Hall, D. B. Goldgof, R. Liu and R. A. Gatenby, "Fine-tuning convolutional deep features for MRI based brain tumor classification.", *Medical Imaging 2017: Computer-Aided Diagnosis,* vol. 10134, p. 101342E, 2017.

[24] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks", *Nature,* vol. 542, pp. 115-118, 2017.

[25] M. Dorfer, R. Kelz and G. Widmer, "Deep Linear Discriminant Analysis", 15 November 2015. [Online]. Available: https://arxiv.org/abs/1511.04707. [Accessed 29 August 2018].

[26] S. Patel, "Chapter 2 : SVM (Support Vector Machine) — Theory", 3 May 2017. [Online]. Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. [Accessed 29 August 2018].

[27] P. Gupta, "Decision Trees in Machine Learning", 17 May 2017. [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052. [Accessed 29 August 2018].

[28] D. Ignatov and A. Ignatov, "Decision Stream: Cultivating Deep Decision Trees", 25 April 2017. [Online]. Available: https://arxiv.org/abs/1704.07657. [Accessed 6 September 2018].

[29] P. Kontschieder, M. Fiterau, A. Criminisi and S. R. Bulò, "Deep Neural Decision Forests", in *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.

[30] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798 - 1828, 2013.

[31] Z.-H. Zhou and J. Feng, "Deep Forest", 14 May 2018. [Online]. Available: https://arxiv.org/abs/1702.08835. [Accessed 1 August 2018].

[32] C.-J. Hsiao, E. Hing and J. Ashman, "Trends in electronic health record system use among office-based physicians: United states, 2007-2012", *Natl Health Stat Report,* vol. 75, pp. 1-18, 2014.

[33] "Outcomes", [Online]. Available: https://www.cihi.ca/en/outcomes. [Accessed 13 November 2018].

[34] J. Ker, L. Wang, J. Rao and T. Lim, "Deep Learning Applications in Medical Image Analysis", *IEEE Access,* pp. 9375 - 9389, 2017.

[35] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical Image Analysis,* vol. 42, pp. 60-88, 2017.

[36] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014.

[37] N. Antropova , B. Q. Huynh and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets", *Medical Physics,* vol. 44, no. 10, pp. 5162-5171, 2017.

[38] R. Paul, S. H. Hawkins and Y. Balagurunathan, "Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma", *Tomography,* pp. 388-395, 2016.

[39] X. Wang, W. Yang, J. Weinreb, J. Han, Q. Li, X. Kong , Y. Yan, Z. Ke, B. Luo, T. Liu and L. Wang, "Searching for prostate cancer by fully automated magnetic resonance imaging classification: Deep learning versus non-deep learning", *Scientific Reports,* vol. 7, no. 1, 2017.

[40] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks", *Radiology,* vol. 284, no. 2, pp. 574-582, 2017.

[41] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot and G. Eramian, "Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network", *Journal of Digital Imaging,* vol. 30, no. 4, pp. 477-486, 2017.

[42] A. Kumar, J. Kim, M. Fulham and D. Feng, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification", *IEEE Journal of Biomedical and Health Informatics,* vol. 21, no. 1, pp. 31-40, 2017.

[43] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz and M. A. Mazurowski, "Deep Learning for identifying radiogenomic associations in breast cancer", 29 November 2017. [Online]. Available: https://arxiv.org/abs/1711.11097. [Accessed 7 August 2018].

[44] Z. Zhu, M. R. Harowicz, J. Zhang, A. Saha, L. J. Grimm, E. Shelley Hwang and M. A. Mazurowski, "Deep learning analysis of breast MRIs for prediction of occult invasive disease in ductal carcinoma in situ", 28 November 2017. [Online]. Available: https://arxiv.org/abs/1711.10577. [Accessed 7 August 2018].

[45] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 4 September 2014. [Online]. Available: https://arxiv.org/abs/1409.1556. [Accessed 7 August 2018].

[47] Z. Li, Y. Wang, J. Yu, Y. Guo and W. Cao, "Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma", *Scientific Reports,* vol. 7, no. 1, 2017.

[48] H.-I. Suk, S.-W. Lee and D. Shen, "Deep ensemble learning of sparse regression models for brain disease diagnosis", *Medical Image Analysis,* pp. 101-113, 2017.

[49] S. Khawaldeh, U. Pervaiz, A. Rafiq and R. S. Alkhawaldeh, "Noninvasive Grading of Glioma Tumor Using Magnetic Resonance Imaging with Convolutional Neural Networks", *Applied Sciences,* vol. 8, no. 1, 2018.

[50] G. González, S. Y. Ash, G. V. Sanchez-Ferrero, J. O. Onieva, F. N. Rahaghi, J. C. Ross, A. Díaz, R. S. J. Estépar, G. R. Washko and COPDGene and ECLIPSE investigators, "Disease Staging and Prognosis in Smokers using Deep Learning in Chest Computed

Tomography", *American Journal of Respiratory and Critical Care Medicine,* pp. 1-49, 2017.

[51] A. Canziani, A. Paszke and E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications", 14 April 2017. [Online]. Available: https://arxiv.org/abs/1605.07678. [Accessed 8 August 2018].

[52] M. C. Chen, R. L. Ball, L. Yang, N. Moradzadeh, B. E. Chapman, D. B. Larson, C. P. Langlotz, T. J. Amrhein and M. P. Lungren, "Deep Learning to Classify Radiology Free-Text Reports", *Radiology,* vol. 286, no. 3, pp. 845-852, 2018.

[53] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[54] T. Clark, A. Wong, M. A. Haider and F. Khalvati, "Fully deep convolutional neural networks for segmentation of the prostate gland in diffusion-weighted MR images", in *International Conference Image Analysis and Recognition*, 2017.

[55] R. McKinley, R. Wepfer, T. Gundersen, F. Wagner, A. Chan, R. Wiest and M. Reyes , "Nabla-net: A Deep Dag-Like Convolutional Architecture for Biomedical Image Segmentation", Springer, 2017.

[56] Q. Zhang , Z. Cui, X. Niu, S. Geng and Y. Qiao, "Image Segmentation with Pyramid Dilated Convolution Based on ResNet and U-Net", 2017.

[57] F. Milletari, N. Navab and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation", 15 June 2016. [Online]. Available: https://arxiv.org/abs/1606.04797. [Accessed 8 August 2018].

[58] C. M. Deniz, S. Xiang, S. Hallyburton, A. Welbeck, S. Honig, K. Cho and G. Chang , "Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Network", 20 April 2017. [Online]. Available: https://arxiv.org/abs/1704.06176. [Accessed 8 August 2018].

[59] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation", 21 June 2016. [Online]. Available: https://arxiv.org/abs/1606.06650. [Accessed 8 August 2018].

[60] L. Shen and T. Anderson, "Multimodal Brain MRI Tumor Segmentation via Convolutional Neural Networks", 2017. [Online]. Available: http://cs231n.stanford.edu/reports/2017/pdfs/512.pdf. [Accessed 8 August 2018].

[61] S. B. Lo, S. A. Lou, J. S. Lin, M. T. Freedman, M. V. Chien and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection", *IEEE Transactions on Medical Imaging,* vol. 14, no. 4, pp. 711-718, 1995.

[62] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, "Scalable object detection using deep neural networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[63] C. Szegedy, S. Reed, D. Erhan, D. Angueloc and S. Ioffe, "Scalable, High-Quality Object Detection", 3 December 2014. [Online]. Available: https://arxiv.org/abs/1412.1441. [Accessed 26 November 2018].

[64] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei and K. Cha, "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography", *Medical Physics,* vol. 43, no. 12, pp. 6654-6666, 2016.

[65] 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.

[66] C. Szegedy , S. Ioffe, V. Vanhoucke and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", in *AAAI Conference on Artificial Intelligence*, 2017.

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

[68] S. Ren, K. He, R. Girshich and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", 4 June 2015. [Online]. Available: https://arxiv.org/abs/1506.01497. [Accessed 9 August 2018].

[69] A. Qayyum, S. M. Anwar, M. Awais and M. Majid, "Medical image retrieval using deep convolutional neural network", *Neurocomputing,* vol. 266, pp. 8-20, 2017.

[70] L. S. Chow and R. Paramesran, "Review of medical image quality assessment", *Biomedical Signal Processing and Control,* vol. 27, pp. 145-154, 2016.

[71] L. Wu, J.-Z. Cheng, S. Li, B. Lei and T. Wang, "FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks", *IEEE Transactions on Cybernetics,* vol. 45, no. 5, pp. 1336-1349, 2017.

[72] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming , K. Gin , J. Swift, R. Rohling and P. Abolmaesumi, "Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four Chamber View", *IEEE Trans Med Imaging,* vol. 36, no. 6, pp. 1221-1230, 2017.

[73] A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*, 2 ed., SAGE Publications Inc, 1990.

[74] P. Mayring, "Qualitative Content Analysis", *Forum: Qualitative Socal Research,* vol. 1, no. 2, 2000.

[75] A. Bryman, *Social Research Methods*, 4th ed., Oxford: Oxford University Press, 2012.

[76] D. Silverman, *Doing Qualitative Research*, 3 ed., London: SAGE Publications, 2010.

[77] A. J. Onwuegbuzie and R. Frels, "Chapter 3: Methodology of the Literature Review", in *Seven Steps to a Comprehensive Literature Review: A multimodal and Cultural Approach*, SAGE Publications Ltd, 2016, pp. 48-64.

[78] "Classification: ROC and AUC", Google Developers, 1 October 2018. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. [Accessed 26 November 2018].

[79] A. Rosebrock, "Intersection over Union (IoU) for object detection", pyimagesearch, 7 November 2016. [Online]. Available: https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/. [Accessed 26 November 2018].

[80] J. Hui, "mAP (mean Average Precision) for Object Detection", Medium, 7 March 2018. [Online]. Available: https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173. [Accessed 26 November 2018].

[81] "Assembly language", Computer Hope, 04 November 2017. [Online]. Available: https://www.computerhope.com/jargon/a/al.htm. [Accessed 11 November 2018].

[82] J. C. Wong, "Greyball: how Uber used secret software to dodge the law", 4 March 2017. [Online]. Available: https://www.theguardian.com/technology/2017/mar/03/uber-secret-program-greyball-resignation-ed-baker. [Accessed 4 November 2018].

[83] B. Hulac, "Volkswagen Uses Software to Fool EPA Pollution Tests", 21 September 2015. [Online]. Available: https://www.scientificamerican.com/article/volkswagen-uses-software-to-fool-epa-pollution-tests/. [Accessed 4 November 2018].

[84] P. Hannon, "Researchers say use of artificial intelligence in medicine raises ethical questions", 14 March 2018. [Online]. Available: http://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A%20NewsFromStanfordsSchoolOfMedicine%20%28News%20from%20Stanford%27s%20School%20of%. [Accessed 4 November 2018].