Imperial College of Science, Technology and Medicine
Department of Computing

Exploring Variability in Medical Imaging

Elissavet (Elisa) Chotzoglou

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy and Diploma
of
Imperial College London

June 2021

Abstract

Although recent successes of deep learning and novel machine learning techniques improved the performance of classification and (anomaly) detection in computer vision problems, the application of these methods in medical imaging pipeline remains a very challenging task. One of the main reasons for this is the amount of *variability* that is encountered and encapsulated in human anatomy and subsequently reflected in medical images. This fundamental factor impacts most stages in modern medical imaging processing pipelines.

Variability of human anatomy makes it virtually impossible to build large datasets for each disease with labels and annotation for fully supervised machine learning. An efficient way to cope with this is to try and learn only from normal samples. Such data is much easier to collect. A case study of such an automatic anomaly detection system based on normative learning is presented in this work. We present a framework for detecting fetal cardiac anomalies during ultrasound screening using generative models, which are trained only utilising normal/healthy subjects.

However, despite the significant improvement in automatic abnormality detection systems, clinical routine continues to rely exclusively on the contribution of overburdened medical experts to diagnosis and localise abnormalities. Integrating human expert knowledge into the medical imaging processing pipeline entails uncertainty which is mainly correlated with inter-observer *variability*. From the perspective of building an automated medical imaging system, it is still an open issue, to what extent this kind of *variability* and the resulting uncertainty are introduced during the training of a model and how it affects the final performance of the task. Consequently, it is very important to explore the effect of inter-observer variability both, on the reliable estimation of model's uncertainty, as well as on the model's performance in a specific machine learning task. A thorough investigation of this issue is presented in this work by leveraging automated estimates for machine learning model uncertainty, inter-observer variability and segmentation task performance in lung CT scan images.

Finally, a presentation of an overview of the existing anomaly detection methods in medical imaging was attempted. This state-of-the-art survey includes both conventional pattern recognition methods and deep learning based methods. It is one of the first literature surveys attempted in the specific research area.

Acknowledgements

Firstly, I would like to thank my advisor, Dr. Bernhard Kainz, for the opportunity he gave me to pursue this PhD in the research field of Medical Imaging and for his encouragement.

I would like to express my deepest appreciation to Prof. Daniel Rueckert for his kind support during the difficult moments in my research studies. Furthermore, I would like to express my sincere gratitude to Prof. Sophia Drossopoulou and Dr. Amani El-Kholy for their support.

I would also like to thank my friends, both in London and in Thessaloniki, for their support and friendship throughout my doctoral studies.

I am also extremely grateful to my family for their patience, support and encouragement throughout this journey.

Especially, this work is dedicated to the memory of my father who, as a surgeon, served medical science selflessly, but suddenly passed away before the completion of this work.

In memory of my father

and

To my dearest mother

Statement of Originality

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

Elissavet Chotzoglou

Acronyms

AD Anomaly Detection

AE Autoencoder

DAE Denoising Autoencoder

VAE Variational Autoencoder

cVAE Conditional Variational Autoencoder

CAE Convolutional Autoencoder

AAE Adversarial Autoencoder

GMVAE Gaussian Mixture Variational Autoencoder

DNN Deep Neural Network

NN Neural Network

CNN Convolutional Neural Network

GAN Generative Adversarial Network

OC One-Class

IF Isolation Forest

DBN Deep Belief Network

RBM Restricted Boltzmann Machine

SVM Support Vector Machine

MIL Multiple Instance Learning

LDA Linear Discriminant Analysis

MRF Markov Random Field

CRF Conditional Random Field

MLP Multi-layer Perceptron

MRI Magnetic Resonance Imaging

fMRI functional Magnetic Resonance Imaging

CT Computed Tomography

SD OCT Spectral Domain Optical Coherence Tomography

PET Positron Emission Tomography

DW Diffusion Weighted

GMM Gaussian Mixture Model

ROI Region of Interest

KDE Kernel Density Estimation

k-NN k-nearest neighbors

MS Multiple Sclerosis

KL Kullback-Leibler

ML Machine learning

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

The main concepts explored in this thesis are automatic *anomaly detection* and *uncertainty* which originates at the human-in-the-loop approach in medical imaging. Both concepts are examined under the prism of *variability*. *Variability* is encountered and encapsulated both in the human anatomy and subsequently is shown in medical imaging as well as in the different stages of medical imaging processing pipelines. An abstract visualisation of the main stages in this pipeline is shown in Figure 1.1.

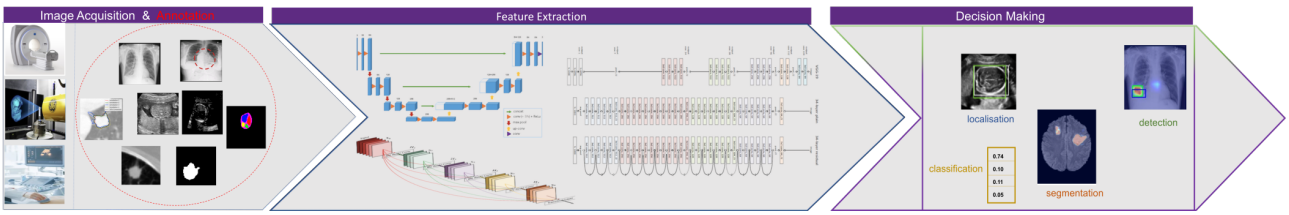


Figure 1.1: An overview of the basic stages in the Medical Imaging processing pipeline. Images in this Figure are borrowed from the following references: [KLN+17, HM16, KCL21, BKM+17a, ÇAL+16, THC+19, HZRS16, HLW16, WPL+17, RS20, AIMB+11]

In human anatomy, it has been recognised that the human body displays a range of morphological patterns and arrangements, often called an anatomical *variation* [Smi21]. There is a plethora of normal variations among individuals within the reference range of normality. Any morphological fluctuation that is beyond the limits of normality is defined as anomaly or malformation. Although in natural sciences *normality* is precisely defined, in human anatomy it

can be considered as a convention [Cha15, ŻSTI+21]. For example, in statistical science, *normality* in a population can be precisely described and modelled by a bell-shaped probability distribution, i.e normal (Gaussian) distribution which is defined by mean and variance. However, in human anatomy, *normality* "is not as precise as one would wish and can be considered an approximation or consensus" [ŻSTI+21]. An example of normal and abnormal variants in the medical imaging field is given in Figure 1.2.

An understanding of the anatomical variations (normal and abnormal) is crucial for performing a range of surgical and medical procedures as well as for the treatment of diseases. For instance, in order to identify abnormalities in fetal anatomy, the key is to understand the range of normal appearances at differing gestations. Furthermore, the identification and accurate detection of anatomic variants that present abnormal manifestations is vital, especially for treating diseases at an early stage. The early detection of a pathological anomaly is beneficial both, for the patient, as well as for the healthcare system. Providing care at the earliest possible stage of the illness increases both the chances for successful treatment and the likelihood of survival. Also, treatment would cause less hassle for the patients, reducing the patient length of stay. At the same time, it enables the healthcare system both to save human and financial resources, as well as to manage and allocate them more efficiently.

Towards this direction, the development of automatic anomaly detection systems can significantly assist experts to make faster and more accurate diagnoses and concurrently reduce significantly the observer dependence and the experts' decision uncertainty. It is therefore highly valued to integrate successfully such systems into the clinical routine.

Substantial advances by deep learning-based methods in various machine learning tasks, make these methods dominant for the implementation of robust and efficient anomaly detection systems. The two key components leading to success of deep learning are: the deep structures of the networks and the use of large annotated datasets [BD20].

However, anatomic variability makes the collection of large datasets that reflect all the anatomical variations (and pathological manifestations) a very difficult, costly, time-consuming and usually impractical process. This leads to limited availability and scarcity of abnormal data (e.g rare disease). Additionally, in many application fields, including medical imaging, it is easier to obtain data that conform to normal behaviour. For instance, many datasets, e.g.

volunteer studies like UK Biobank [PMB+13], consist of images from predominantly healthy subjects with a small proportion of them belonging to abnormal cases.

Consequently, an effective way to treat the abnormality detection problem in medical imaging, is to consider it within a framework of one-class classification [PCCT14]. In this case, one class (i.e normal subjects) is well-sampled while others are (severely) under-sampled. Training models based only on normal samples (or with few abnormal samples), with minimal supervision, enables identifying anomalies/patterns that differ from normality. Additionally, the need for use of annotated data is reduced, since the anomaly is defined implicitly by the appearance of the normal set.

Based on the above facts, building an anomaly detection system consists of two main steps. In the first step the model is trained with only normal (or with few abnormal) data. In the next step, the abnormality is identified based on the deviation from the normal cases. [PCCT14].

A case study of such an automatic anomaly detection system is presented in this work. A framework for detecting fetal cardiac anomalies during ultrasound screening is proposed. The model is trained only using healthy subjects in a unsupervised framework.

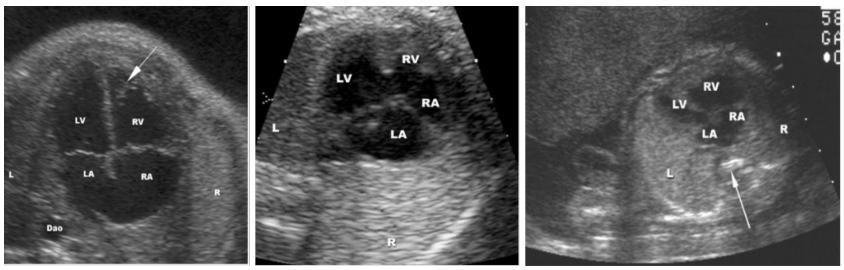


Figure 1.2: From left to right: (A) Transabdominal US image (four-chamber view) shows the moderator band (arrow) in the right ventricle, the left ventricle, the right atrium, the left atrium, and the descending aorta (Dao), which is anterior to the echogenic spine (normal view). (B)Transabdominal US image (four-chamber view) shows absence of the interventricular and interatrial septa, thus producing connections between the ventricles and between the atria (Endocardial cushion defect) (C) Transabdominal US image (four-chamber view) shows that the left ventricle is small relative to the right ventricle and the left atrium is small relative to the right atrium (arrow=spine) (Hypoplastic left heart syndrome) [BDGA02]

.

Despite the significant improvement in automatic anomaly detection systems, clinical routine continues to rely heavily on the contribution of medical experts to diagnose and localise abnormalities. Integrating human expert knowledge into the medical imaging processing pipeline entails uncertainty, which is mainly quantified from inter-observer *variability*. Specifically, *variability* is sourced from the subjectivity in the perception of the boundaries of an anomaly, such as a lung tumor, among human experts (e.g radiologists). This subjectivity in various scenarios (e.g. medical experts from different medical centers) causes disagreement over the annotations of a pathology in medical images. Examples of inter-observer variability are given in Figure 1.3. Delineation agreement has been assessed by computing the geometric agreement between contour delineations (or comparing contours to the gold-standard), utilising metrics such as the Dice similarity coefficient, the Hausdorff distance or the mean surface distance [VJMH16, ANS19]. Also, several other methods for quantifying inter-rater variability have been applied, such as Cohen's Kappa and its variations, Pearson correlation coefficient, Bland-Atman plots and the intra-class correlation coefficient [RPA17, AHS18]. In some cases the variations in delineating the structures of interest, like breast tumors (and organs at risk structures), even by well-experienced observers from different institutions, are substantial [LTA$^+$09]. Inconsistencies in delineating target and organ-at-risk volumes and structures have been identified, for several tumor sites. The largest variation for delineation of target volume, based on the ratio of the largest to the smallest delineated volume (V_{max}/V_{min}) was reported for oesophageal tumours, head and neck, lung cancer, Hodgkin's lymphoma and sarcoma [SP16]. As it is mentioned in a breast-cancer radiotherapy study [LTA$^+$09], the overlap in the structure of interest was almost 10%, while volume variations show standard deviations of up to 60%. In another study [RST$^+$11], the median standard deviation in (gross) tumor volumes among experts was about 6% of the average volume in soft-tissue sarcoma. These errors often lead to significant variations in dosimetric planning for various types of radiotherapy. For instance, in radiation oncology, radiation could miss a tumour part, while at the same time it might cause damage to healthy/normal tissue.

Additionally, common (observer's) factors that are related with this *disagreement* include diseases complexity, physicians lack of experience and fatigue, imperfect information and the quality of the available medical images. For instance, it has been reported that less experienced physicians contoured larger tumor volumes than experts [Nje08].

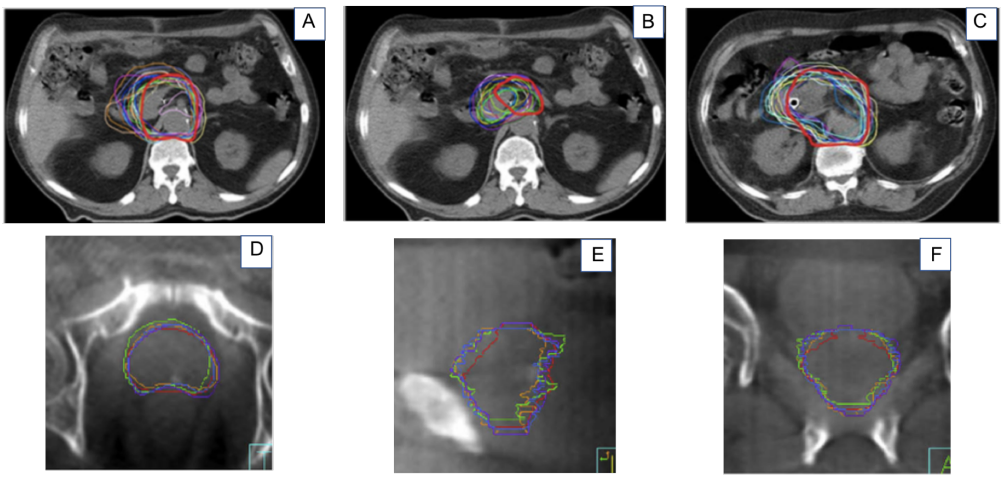


Figure 1.3: Top row: A-B-C: Representation on axial planes (CT scan for detecting pancreatic adenocarcinoma) of interobserver variation between 18 centers. Red solid line represents the center of reference [CMM+14]. Bottom row: Delineation of prostate (cone beam computerised tomography images) from five observers [WBJC08]

Furthermore, an accurate delineation process requires adequate hardware and software [AoR21]. For instance, some radiologists find the utilisation of tablets with pens more helpful for faster contouring, reducing at the same time the need for correction of ROIs. A comparison of different systems for input contouring i.e mouse-keyboard or pen-tablet user devices is given in [iOG11]. Based on the study [MBvH21], minor interobserver variations in the delineation process were observed, caused by the difficulty to outline "fuzzy" boundaries in tumors, using the existing contouring tools. Additionally, variability in image quality is also introduced across facilities, i.e across a large cohort of CT scans acquired from different clinical sites [SZE+21]. Variability can be also reported after an upgrade of a system like an MRI system [PKC+19].

An additional factor that causes disagreement in contouring is the fact that, consensus contouring guidelines for target and organ-at-risk volume delineation are often insufficiently used [LLS+20]. Additionally, in some cases, there is a need for a better interpretation and application of the existing protocols amongst experts [MBvH21].

All the above factors affect the quality of diagnosis, which is subsequently followed by inaccurate or improper treatment.

There are some strategies in clinical practice that have been proposed to reduce the uncertainty derived from inter-observer variability. Multi-modal imaging [SP16], introduction of clearer

and comprehensive delineation protocols and guidelines (which should be followed strictly by all members of a department), clinician to clinician peer review for improving contouring, IT solutions for cross-site peer review of contours [oR17], use of atlases and autocontouring tools for organ-at-risk delineation [VMJH16] as well as more training of experts are some of these strategies [VMJH16, SP16]. However, there are still open issues which are related with the implementation of these strategies in the daily clinical practice.

From the perspective of building an automated medical image analysis system, it is still an open issue, to what extent this kind of *variability* and the resulting uncertainty are introduced during the training of a model and how it affects the final performance of the task. As it is reported, most of the annotator's disagreements arise in the more difficult and ambiguous cases [TSS+19]. At the same time, such cases are highly likely to be misclassified by a well-trained deep neural network or to be classified in a class with high model uncertainty [LIO19]. Based on this hypothesis, it is very important to explore the effect of considering the inter-observer variability on the reliable estimation of the model's uncertainty. Therefore, an examination of the potential relationship between these two quantities is needed. Towards this direction, leveraging model's uncertainty, inter-observer variability and task performance is a prerequisite and can yield powerful models, in terms of reliability and patient safety. At a later stage, these models could be successfully embedded in clinical practice.

Finally, intrinsic *variability* is introduced at the first stage of the medical imaging pipeline, i.e. during the process of medical image acquisition. This kind of *variability* can come from a variety of reasons. For example, with ultrasound positioning, there are indications of inter-observer *variability* and potential introduction of errors due to the pressure to the patient's lower abdomen during image acquisition [int16]. However, the most common cause of intrinsic *variability* in medical datasets can be detected in cases where data are acquired by different hospitals, using different machines/scanners and imaging modalities, following different medical protocols in various subject populations. In this scenario, *domain shift* is common, i.e. data coming from different distributions but related domains [YDZ+19a]. To alleviate this issue, many algorithms for domain adaptation have been proposed [YDZ+19a, YDZ+19b, GL21]. Better progress has been achieved for the domain shift caused by different scanners, centers and protocols [YDZ+19a]. Domain adaptation related to different modalities is still a subject

under extensive investigation due to the large domain shift between modalities [YDZ+19a]. However, the examination of this kind of *variability* is out of the scope of this thesis.[1]

1.2 Thesis Outline

In this thesis, a one-class abnormality detection algorithm which utilises only normal samples in the training phase, is presented. Furthermore, the relationship between segmentation performance, segmentation uncertainty and inter-observer variability is examined. Our work is not limited to a specific modality or specific pathology. Application fields include Ultrasound imaging and Computed tomography scans. Furthermore, different variants of pathologies are examined, such as tumors and heart disease which present different characteristics in order to be recognised.

Chapter 2 provides an introduction to **Anomaly Detection** (AD) and **Uncertainty estimation** methods. For Anomaly detection, a brief overview and taxonomy of the methods, as well as of the work before the era of Deep Learning is given. Subsequently, a brief overview of Deep Learning methods along with related works is provided. Finally, focusing exclusively on medical imaging, a study of existing methods in combination with the available Datasets and evaluation protocols is presented. For Uncertainty estimation, the study is mainly focused on uncertainty quantification in Deep Neural networks. In this context, mathematical background and a brief description of existing methods in Medical imaging is given.

In *Chapter 3* a case study of an automatic anomaly detection algorithm is presented. The aim of the algorithm is to detect a subtype of Congenital Heart disease in fetal Ultrasound Imaging using an unsupervised method. A population of normal subjects is utilised to train an algorithm and then the method is tested on different test sets consisting of both normal and abnormal cases. An examination of the advantages between single-frame and multi-frame processing is investigated. Furthermore, the ability of the algorithm to localise the area of pathology is examined. A comparison with other state-of-the art methods is also assessed.

Chapter 4 focuses on inter-expert variability. Having identified an anomaly, in many cases,

[1] Domain adaptation methods in the context of the proposed frameworks (anomaly detection/uncertainty estimation) have not been examined since the prerequisite for this was the availability of similar, appropriate datasets from a different site/modality (target data).

manual annotations, which are carried out by human experts, suffer from inter-rater variability. A framework for analysis and comparison of inter-rater variability with performance and uncertainty in medical imaging segmentation is presented. The application domain is lung tumor segmentation in CT scans. Two state-of-the art methods are applied to a 3D imaging segmentation task. An investigation of the correlation between inter-rater variability, uncertainty estimation and segmentation performance is conducted. A metric for capturing the correlation between inter-rater disagreement and segmentation uncertainty is also presented. Finally, both quantitative as well as qualitative analysis are presented.

Chapter 5 summarises the key scientific achievements of this thesis. Furthermore, a discussion about the limitations of the work is given. Potential new research directions are also suggested which are left as future work.

Chapter 2

Background

In this Chapter, a comprehensive overview of the existing literature in anomaly detection and Uncertainty quantification, focusing mainly in Medical Imaging, is provided. In Section 2.1.1 an introduction to anomaly detection definition, taxonomy and basic concepts is given. Subsequently, in Section 2.1.2, a brief description of the most common deep learning-based networks, which have been used for anomaly detection is presented. Finally, in Section 2.1.3 a survey of medical imaging anomaly detection methods is presented. It consists of the description of the existing methods, the presentation of the available datasets and the evaluation protocols. In the second part of this Chapter, the basic concepts for Uncertainty quantification, as well as an introduction to Bayesian machine learning are presented (Section 2.2.1 and 2.2.2). Different ways for computing uncertainty in deep learning models are also introduced (Section 2.2.4). Finally, our study focuses on uncertainty estimation in medical imaging, presenting the existing works in this research field (Section 2.2.5).

2.1 Anomaly Detection

2.1.1 Introduction

Anomaly detection (AD) is defined as the problem of finding patterns that do not conform to expected (normal) behaviour (i.e outliers) [CBK09, ABA06, HA04]. In [Haw80], Hawkins defines outlier as: “an observation which deviates so much from other observations as to arouse

suspicions that it was generated by a different mechanism". A typical example of anomalies in 2D space [CBK09] is shown in Figure 2.1. As can be seen in this Figure, there are two sets of normal samples S_1 and S_2, and three areas of anomalies O_1, O_2 and O_3. These non-conforming

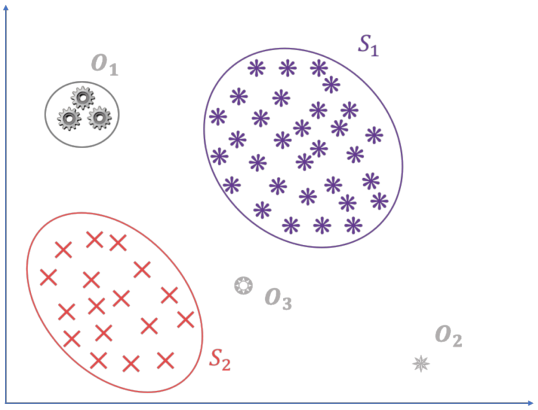


Figure 2.1: A typical example of anomalies induction in a 2-dimensional dataset

patterns are usually referred to as anomaly-(ies), outlier(s), surprises, irregularities, abnormalities, unexpected events, discordant observations. In this thesis, the terms anomalies (or abnormalities) and outlier(s), two of the most often used terms, will be utilised interchangeably. Identification of anomalies could give important and crucial information for each application domain. For instance, it could be either a sign of malicious activity in network (e.g credit card fraud) or a sign for the existence of a pathology in the medical field or finally an indicator for a fault in a factory system.

Anomalies can be classified in three different categories based on their nature (type of anomaly). The first category is, *point anomalies*, where an individual data instance is anomalous with respect to rest of the data [CBK09] (Figure 2.1). Most of the current research is focused on this type of anomaly, which is also considered as the simplest type of anomaly (i.e an anomaly in MRI could be a sign for a brain tumour).

Contextual (or conditional [SWJR07]) *anomalies* is another category where, a data instance is anomalous in a predefined/pre-specified context and not otherwise. The notion of context should be defined as part of the problem formulation. An individual data instance could be considered as anomalous in a specific context and as a normal in a different context formulation. Each data point is defined utilising two sets of attributes: (a) contextual attributes and (b) behavioral attributes. The former type of attributes is used to define the context for the

data instance, while the latter one defines the non-contextual information for the data sample. A typical example of *contextual* anomaly detection problem, is the discovery of anomalies in monthly temperature data from temperature time-series data. In order to define an anomaly as contextual, the availability of contextual attributes is necessary. An example of contextual anomaly detection in healthcare is a framework utilised to detect (medical) utilization instances that are unexpected given patients' clinical characteristics. In this framework, contextual attributes could be the patient characteristics (e.g comorbidities) and behavioural attributes could be the number of clinical visits [HWS+12].

Finally in the last category, *Collective anomalies*, a collection of data instances is anomalous compared to the rest of the data stream. The individual data instances in this category may not be anomalies by themselves, but only in the co-existence with other data can be characterised as anomalous. An anomaly could fall in this category, only if the data samples are related. An example that belongs to this category is the examination of a human electrocardiogram, where a different pattern of cardiac rhythm for a not normal duration, compared to the rest part of the electrocardiogram, could be characterised as an outlier [CBK09]. A value of each individual sample does not contain information on whether the whole electrocardiogram signal is anomalous.

Point and collective anomalies could be also examined as contextual anomalies, in the cases that contextual information exists and is utilised in the detection problem formulation.

The study of AD begun in the early 19^{th} century [Edg87] and up to date many different techniques have been developed. However, in order to resolve the AD problem, many factors should be taken into consideration, which mainly originate from the nature and quality of data, the lack and scarcity of annotated data and the type of anomaly which may not be the same among the application fields.

One of the most important aspects in AD, is the availability of well annotated datasets. The level of annotation differs from one dataset to another and based on the type of annotation (level of supervision), the AD methods can be divided in three different categories. In the first category, *supervised anomaly detection*, data for both classes, normal and abnormal, together with labels/annotations is available. However, in most cases, anomalous data are much fewer compared to normal ones. This happens because building a well annotated dataset is a time-

consuming, costly, process and requires a lot of human effort, especially in some special research fields, such as medical imaging. Supervised anomaly detection is similar to building a predictive model for normal and anomalous class.

In *Semi-supervised anomaly detection* the basic assumption is the existence of both unlabeled and labeled data, where most of the data are unlabeled and only few labeled data samples are available [RKV+21].

Finally, in *Unsupervised anomaly detection* there is no available information for the labels of the data [RKV+21].

The output of an anomaly detection algorithm could be either an anomaly score or a label for normal/abnormal data. In the former case, an anomaly score is derived for each data sample during the test phase and based on this score, a sample is characterised as a normal or abnormal. In order to decide, a cut-off threshold usually is computed. Alternatively, a top-k list of samples based on the score is derived. In the latter case, a label (normal/abnormal) which indicates the class where each sample belongs to, can be derived either based on the score and threshold or directly as the output of a (classification) model.

Input data can be categorised based on existing relationship between data instances (nature of data) [CBK09]. Input data could be categorical (e.g fraud detection), discrete sequences (e.g bio-informatics), spatio-temporal (e.g climate), time-series (e.g healthcare), spatial (e.g vehicular traffic data), or graph data (e.g social networks, epidemiology).

AD methods can be generally categorised into the following categories: (a) probabilistic methods, (b) Distance-based methods, (c) Reconstruction-based, (d) classification-based methods and (e) Information-theoretic based approaches [PCCT14, CBK09].

A figure which shows the taxonomy of AD based on the type of anomaly, method and level of supervision is shown in Figure 2.2.

Probabilistic approaches are based on the estimation of the probability density function (pdf) of (training) data samples. A threshold is set into the resultant distribution in order to define the boundaries of normality in the data space. There are many techniques which lie in this research area and present differences regarding their complexity [PCCT14]. The simplest statistical

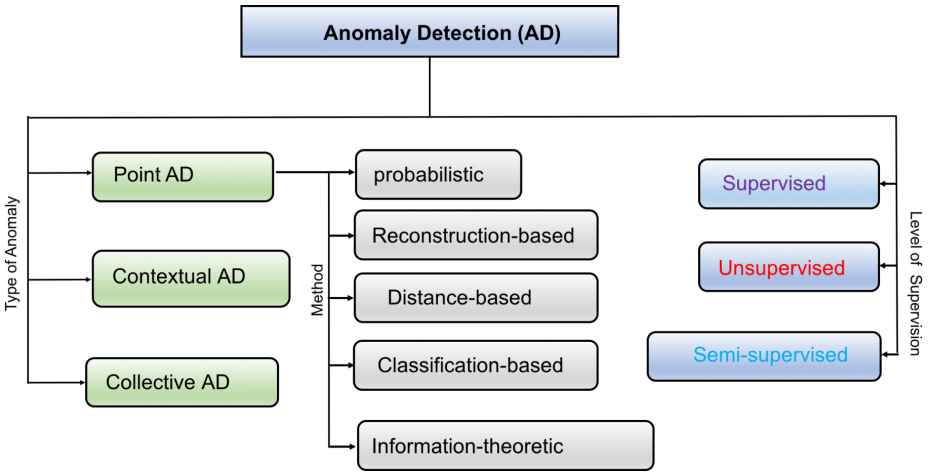


Figure 2.2: A taxonomy of AD based on the type of anomaly, method and level of supervision.

techniques are based on statistical hypothesis tests such as Grubbs' test [Gru69] and box-plot rule [SL05]. More advanced probabilistic techniques (estimation of data density) can be divided into parametric and non-parametric methods.

Parametric approaches assume that data are generated from a parametric distribution with specific parameters. The most commonly used distribution is the Gaussian distribution and its parameters are estimated using Maximum Likelihood Estimation (MLE). More complex forms of probability distributions can be modelled by mixture models such as Gaussian Mixture Models (GMMs) [RT94] [Bis93]. State space models have been also used in anomaly detection mainly in time-series. Two of the most common methods include Hidden Markov Model (HMM) [YD03] [NPF11] [Smy94] and Kalman filter [GH00] [QW07].

Opposite to parametric methods, non-parametric approaches do not assume a fixed model structure. The simplest non-parametric methods are the histogram-based methods which are used to maintain a profile of normal data. Kernel Density Estimator (KDE) [Hä90] [Par62] is also a popular non-parametric approach where the probability density function is estimated using many kernels distributed over the space of data. Although non-parametric estimators perform well for low-dimensional problems, they suffer from the curse of dimensionality [RKV+21]. Recent proposed deep learning-based methods such as energy based models (Deep Belief Networks [HOT06]), Variational Autoencoders [KSW15] and Generative Adversarial Networks [GPAM+14], overcome, in most cases, the above issue. A more analytical description of these methods will be given in Section 2.1.2.

Distance based approaches are based on well-defined distance metrics to compute the distance/similarity between two data points [PCCT14]. The basic assumption in this category, is that normal data have dense neighbours while abnormal instances are far apart from their neighbours [CBK09]. Distance based methods include Nearest neighbour and clustering approaches. The concept of nearest neighbor method is widely applied in AD problems. They require a distance/similarity measure for computing the distance between two data samples. Euclidean distance or Mahalanobis distance are examples of such measures. Nearest neighbours methods could further split into two main groups: the first one consists of methods that utilise as anomaly score the distance of a data point to its k^{th} nearest neighbor [HKF04]. The second type of methods obtain the relative density (of the neighborhood) of each data point in order to compute its anomaly score. Local Outlier Factor (LOF) [BKNS00] algorithm and its variants [CF03] [TCFC02] belong to this category. The main advantage of this type of algorithms, is that they are data driven and no assumption about the underlying distribution is needed. Also, it is very easy to adapt the above methods for any kind of data. However, the computational complexity as well as the definition of efficient distance measures, especially in complex data, are the main challenges.

In clustering based approaches, the basic assumption is that normal data belong to large and dense clusters and anomalous data do not belong to clusters (or belong to very small clusters). Algorithms that belong to this category are k-means Clustering [SZ05], fuzzy c-means [Bez81], probabilistic c-means [KK93] and Expectation Maximization (EM) based methods [BB04]. In terms of complexity, clustering based approaches can have quadratic complexity, although its complexity is highly related to the algorithm that is used to generate clusters. Another disadvantage is that many clustering-based algorithms are effective only in the cases that anomalies do not form clusters among themselves [CBK09].

Reconstruction methods include neural networks and subspace-based approaches as presented in [PCCT14] review of novelty detection. Training models using reconstruction objective is one of the most common approaches in AD research area. Models are optimised to reconstruct well normal instances using a reconstruction objective during the training phase. In the test phase, a data instance is flagged as an anomalous in case the model fails to reconstruct accurately the specific instance. Most recently, deep learning based methods were utilised for learning

reconstruction basis function. Such methods include autoencoders and generative adversarial networks. Since this kind of methods show great performance, within the deep learning framework, a more detailed description will be given in the next Section 2.1.2. Another type of reconstruction-based method is subspace methods (or spectral methods [CBK09]). In this kind of methods the basic assumption is that data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different [CBK09]. Several techniques use Principal Component Analysis (PCA) [Jol02, Haw74] for projecting data to a lower dimensional space. Variants of PCA such as Kernel PCA [SSM98, Hof07], Bayesian PCA [Bis99], Robust PCA [Kwa08], Probabilistic PCA [TB99] have also been attempted in the context of the AD framework.

Classification based anomaly detection methods are used to train a model/classifier using normal and abnormal data from a labelled dataset. In the testing phase, a new data point is classified into a class based on the training model. Classification models must be able to handle also rare classes and imbalanced datasets. Neural network-based approaches is the most representative type of algorithms for novelty detection in this category, especially for the multi-class setting. Examples of neural networks based methods include multi-layer perceptrons (MLPs) [AF02, SM04, CBD$^+$90], Hopfield networks [CMHN02, Jag91] and Radial Basis functions (RBFs) [JS02, Bis93], auto-associative networks [DH02a, Aey91]. Bayesian networks and its variants have been also proposed for multi-class AD [BWJ01, DH02b, WMCW03]. The networks compute the posterior probability of a class label given a test data instance.

Support Vector Machines [Vap95] have been widely used for two class or multi-class classification, as well as for detection of abnormalities. For instance, Robust Support Vector Machines (RSVMs) [HLV03] have been applied in intrusion detection. For the latter case, i.e one-class classification, One-class SVM (OCSVM) [SWS$^+$00], [MY02] has been successfully applied. One-class Support Vector Machines (OCSVMs) is proposed by [SWS$^+$00] and its aim is to define a boundary in the feature space, by separating the training data from the origin in the feature space with maximum margin. Examples of application of OCSVMs include [MY02] [JSR$^+$20] [EES06]. To overcome some of the OCSVMs shortcomings, the one-class kernel Fisher discriminant classifier was proposed by Roth [Rot04]. This method relates kernelized one-class classification to Gaussian density estimation in the induced feature space. Another

popular approach is Support Vector Data Description [TD04] algorithms. Support Vector Data Description (SVDD) [TD04] defines the novelty boundary as being the hypersphere with minimum volume that encloses all the normal training samples. Then an anomaly is assessed by determining whether the test data point lies outside the hypersphere. Extensions to SVDD have been proposed [LTMS10, LTMS11, LLC10]. The main drawback of these methods is the complexity which is associated with the kernel functions. Recent deep learning approaches, such as Deep SVDD algorithm [RVG+18] and deep OC-SVM [ERKL16] attempted to alleviate the above issue. [1]

Finally, the last category of AD methods is *Information theoretic* techniques. The basic assumption is that *anomalies in data induce irregularities in the information content of the dataset* [CBK09]. An information theoretic measure is required in order to detect anomalies. There is no need for distribution assumption for the data. It can be applied in an unsupervised way. However, it has exponential computational complexity, although approximate techniques have been proposed that have linear time complexity [CBK09]. It is difficult to assign an anomaly score using information-theoretic approaches for a test sample. The selection of information theoretic measure is also a challenge for the application of these methods in AD. Kullback-Leibler divergence is one of the most popular information-theoretic metrics utilized in novelty detection frameworks [Gam06, FS10, IB09].

Anomaly detection is applied in different application domains such as medical imaging, fraud detection, industrial damage detection, sensor networks, speech recognition, fault detection in web applications, traffic monitoring, video surveillance, etc. Table 2.1 presents an indicative list of references based on the application field.

[1] Based on the novelty detection review [PCCT14], OCSVM and SVDD algorithms (and their variants) formed *Domain* based approaches since they require a boundary to be created based on the structure of the training data. They describe the target class boundary or the *domain* and not the class density.

Application domain	**References**
Fraud detection	[AFR97], [GR94], [BLH99], [JDGSSC97], [BH99], [FP99], [CEWB97], [Aga07], [PAL04], [BFD+96], [THHT98], [CC10], [TL11], [AMI16], [ZS15], [SSB+17], [WKR+17], [RH18], [KZ17], [ZYW+19], [PY18], [SHK18], [FSP+19], [CJ18], [AGAPN18], [JGZ+18], [HW17], [AGT16], [AAK18], [Abr18], [LYL+18], [CEH18], [FCTZ16], [Lu17]
Medical/Biological Anomaly detection	[HFLP01], [LuK00], [Rob02], [SWYT03], [CB00], [WMCW02], [WMCW03], [LKFH05], [CKPB18], [SHN+18], [TPMO14], [WGM+11], [ZWBC16], [LJP16], [LX18], [LLDL20], [DWH18], [CHHC20], [CLT+18], [KY18], [ZYC+19], [MLY17], [CK18], [ZGC+20], [LURQ18], [SESG+18], [EKN+17], [WZX+16], [IGV+18], [SSW+17], [SSW+19], [SPS01], [THCB95], [OYU+19], [ZCLZ16], [SSL+17]
Network Intrusion detection	[JZK03], [HBH+16], [ADE20], [LL19], [YC02], [EAP+02], [WD01], [NC03], [SCSC03], [LCD05], [SZ02], [GWC98], [YTWM04], [SOS02], [BDD+01], [KMRV03], [DVF+98], [GAS05], [ZJM+01], [PN97]
Fault/Damage/Industrial detection	[GMEK99], [KLLH07], [DJC98], [YKH01], [RS97], [DH02a], [JS02], [PMD+95], [Man02], [SWF01], [IYC+17], [FHN+16], [BMHK+17], [FYKP17]
Text Anomaly detection	[BHMY99], [FP99], [MY00], [MY02], [ACD+98], [SZ05], [Sri06], [FG20], [MSS12], [KWAP17], [Sd16]
Speech/Audio/Sound Anomaly detection	[Yan20], [RN19], [KTE+19], [KSU+19] [NPF11], [FPS+16], [VGT+07], [COL+13], [ABK+00]
Sensor Networks	[JKR06], [BSG+06], [VPHC+06], [CPGM06], [ZSGL07], [PS15], [BIT+17], [CL19], [OGIR14]
Social Networks/Web Applications	[SQCF05], [IK04], [SHL17], [CPS17], [LC17], [AKA17], [GKK+19], [ZCY+17]
Video Surveillance	[SCS18], [YMR16], [GLS+16], [KLK+16], [DSDVL02], [MPKM16], [BGS08], [XRY+15], [DH02b]
(IoT) Big Data Anomaly Detection	[LN18], [MK18], [ZGL+18], [MKM18]

Table 2.1: An indicative list of references for some of the most common application domains in AD

2.1.2 Anomaly Detection in the Deep Learning Era

Deep Learning methods have been widely applied to AD applications. Depending on the type of data, sequential or non-sequential, different types of architectures and training algorithms have been tested. A brief introduction to the most common deep learning-based networks, which have been used in AD, is given below.

(Deep) Convolutional Neural Networks (CNNs) [LBBH98, KSH12] is a class of deep neural networks. They have been applied to both images, as well as in sequential data. In an AD framework, CNNs are used both as classifiers and feature extractors. They have been applied to different application domains such as the medical domain [IGV+18], pedestrian detection [PYS+20], fraud detection [AGAPN18, CEH18, Lu17], multi-variate time series [WZX+16].

Long short-term memory (LSTM) [HS97, GJ14] is a type of Recurrent Neural Networks (RNNs) [WZ95] applied mainly to sequential data. It allows the network to retain long-term dependencies between data at a given time from many previous timesteps [NPTTH20]. LSTMs are applied to different domains from aircraft data [NS16] to video anomaly detection [MS16, LLG17] and IoT anomaly detection [ZGL+18].

Autoencoders [SH06, HZ94] represent data within multiple hidden layers by reconstructing the input data, effectively learning an identity function. Convolutional autoencoders [MMCS11] have been widely applied to image AD while autoencoders based on LSTMs [SK19] have been applied to sequential data. The autoencoder can be used for AD mainly in two different ways. In the first case, the key idea is that normal data should be reconstructed accurately, while abnormal samples should fail to reconstruct and thus they will derive larger reconstruction error. In the second case, autoencoders are a key element in a AD framework, as a feature extractor, combined with other deep learning architectures such as CNNs, LSTMs or gated-recurrent units [CvMG+14] (to form Gated recurrent unit autoencoders) [APCC19, GEY18, CMC17, GLL+19].

Furthermore, *Variational autoencoders* (VAEs) [KW14, RMW14] are (deep) latent variable models which consist of an encoder (inference model) and a decoder network (generative model). A VAE is trained by maximising the evidence lower bound (ELBO). Many works for anomaly detection based on VAEs have been presented, such as [PBC+19, LX18, BDW+20, ZKP+18, USHE19]. Adversarial autoencoders [MSJ+16] have also been applied successfully in AD [RSNS19, SKFA18]. The adversarial autoencoder is a probabilistic autoencoder that uses an adversarial training procedure (generative adversarial network) to perform variational inference.

Restricted Boltzmann Machines (RBMs) [SMH07] is a generative probabilistic graphical model which can be considered as a stochastic model. RBMs after the training procedure provide a closed-form representation of the distribution underlying the observations [LZG15, FI12]. Consequently, it can be used to compare the probabilities of (unseen) observations and to sample from the learnt distribution [FI12]. RBMs are widely applied as a building block of a multi-layer learning architecture called Deep Belief Networks (DBNs) [HOT06]. The idea is that the hidden neurons extract relevant features from the observations. These features can serve as input to another RBM. There are a few works where DBNs have been utilised for AD [TPMO14, WGM+11, ZWBC16, MVDM17].

Generative Adversarial Networks (GANs) [GPAM+14] have shown great performance in the AD problem. GANs are generative models which consist of two different networks, namely a generator and a discriminator model. The two networks are trained simultaneously in an adversarial way. The generator is trying to capture the data distribution while the discriminator tries to distinguish between real and synthetic data (data which come from the generator). GANs-based anomaly detection frameworks are shown to be effective in different application domains [DVR+19, FN19, SSW+19, AAAB19].

Similarly to the conventional methods, deep AD methods have been applied to many different application fields, such as fraud detection, intrusive detection, image processing, medical/healthcare analysis, speech recognition, video surveillance, etc. Examples of deep learning anomaly detection methods with application to the above domains are included in the Table 2.1.

2.1.3 Anomaly Detection in Medical Imaging

The detection of anomalies in Medical Imaging is one of the most important components in the medical image analysis pipeline. Abnormality detection (AD) methods in medical imaging try to mimic the way expert clinicians use to recognise anomalies in medical images. Experts are familiar with normal patterns of anatomy, being also aware of the healthy characteristics, such as shape, size, opacity and position. Hence, they can easily recognise abnormality (e.g lesion, tumour), which is present in the image, by comparing it with the normal tissues. AD has been applied to different organs such as brain [BWAN19, SCWZ20, BWAN18], breast [ASF+19, FKM+20, ZLS+19], lungs [USHE19], heart [ACZ+20, GZZ+20], head [SHN+18], retina [ZGC+20, SSW+17, SSW+19, ZXY+20], ovaries [VMT98], prostate [LCW+17, RSNL+19], nasopharyngeal structure [ZLCH03] and also to human skin [LX18, SCWZ20, LLDL20].

A factor that should be taken into consideration in medical imaging AD, is the acquisition modality. Image modality plays an important role in the image analysis pipeline since different imaging modalities could produce different images for the same structure. The most common modalities are Computed tomography (CT), Positron Emission Tomography (PET), X-Rays, mammography, Magnetic Resonance Imaging (MRI) and Ultrasound (US) imaging systems. Different tissues have different densities among the variant imaging systems. Depending on the medical imaging modality, abnormalities could appear with different brightness [TCSHPFR09].

The first work attempted to detect and localise anomalies was [BS73] in 1973, for localisation of tumours in radiographs. In this work, authors using pattern recognition methods try to detect tumors in the chest and the liver using X-rays and radioisotope scans, respectively. The input image was digitised and then image enhancement techniques were applied in order to obtain a good representation of the edges and directional contour information of the initial image. Then, using dynamic programming methods, smooth closed curves were detected and finally, using linear functional, the closed curves which were more representative of the tumor edges were selected.

Through the years, algorithms belonging to almost all the various categories, as they were presented in the previous Section 2.1, have been applied for the detection of abnormalities in

medical imaging.

A wide variety of methods have been applied in medical imaging AD such as Markov Random Field(MRF) [Ger03] and Conditional Random Field (CRF) [LSM+05] methods. In [Ger03] the brain tumor segmentation task is considered as an AD problem. Training is performed exclusively on healthy subjects and anomalies are detected based on the principle of the deviation from the normalcy. The method extends the Expectation Maximisation (EM)-based approach with shape descriptors and a novel multi-level MRF approach is derived. In [LSM+05], Conditional Random Fields (CRFs) together with Support Vector machines, formed the Support Vector Random Fields (SVRFs) method, which was proposed for anomaly segmentation of brain tumours in T1 and T2 MRI brain images. SVRFs are derived as Discriminative Random Field (multi-dimensional extension of CRF) which takes advantage of the generalisation properties of SVMs. In another work [ZCM+99] authors proposed a discrete wavelet transform (DWT)-based multiresolution MRF to detect tumors in mammograms. More precisely, in the preprocessing stage, Wavelet decomposition was utilised for image decomposition and its output was used as input to fractal analysis in order to compute the roughness of each and every pixel in the image. Then the fractally analysed image was used as input to a dogs-and-rabbits clustering algorithm and the clustered image was used to initialise MRF segmentation. Finally, following the application of merging techniques in order to group the pixels after segmentation, a binary decision tree was used to classify regions as abnormal. Gaussian Mixture Models (GMMs) have also been applied to multiple sclerosis lesion detection and segmentation on brain MRI [FGG09]. Authors proposed a Constrained Gaussian Mixture Model (CGMM) which is based on a mixture of multiple spatially oriented Gaussians per tissue. In another work [WMAG07] GMMs have been applied in multiple sclerosis lesion detection using multichannel data including “fast fluid attenuated inversion recovery”(fast FLAIR or FF).

Non-parametric probabilistic methods have also been proposed. Local Component Analysis, which is an extension of Parzen windows, was applied in neuroimaging datasets for outlier detection [FVPT12]. Furthermore, a histogram based method was proposed in [NJW14], where the method was based on applying an enhanced gravitational optimization algorithm on brain histogram analysis results for multiple sclerosis lesion detection.

Kernel density estimation (KDE) combined with deep neural networks is attempted in [ECKK21] for the out-of distribution detection in MRI datasets. The feature probability density functions (pdfs) of each channel in a pre-trained DNN is estimated using KDE on an In-distribution training dataset. The method is applied in classification and segmentation task.

Distance based methods have been utilised including both Nearest neighbour-based and clustering-based approaches. In [WWILT+06], statistical k-nearest neighbor(k-NN) was utilised, combined with template-driven segmentation, for automatic segmentation and detection of multiple sclerosis lesion subtypes with multichannel MRI. In [YFH+13] the Naive Bayes Nearest Neighbor (NBNN) method was proposed for liver lesion detection in CT images. In [ZLCH03], authors utilised a 4-stage system to segment nasopharyngeal carcinoma, a malignant skull base tumor, on MRI data. In the first stage, an initial head mask region is generated in order to remove the noise from the image. Semi-supervised fuzzy c-means, a knowledge-based image analysis procedure, is applied for segmentation. In the third stage, distance transform and morphing is applied to perform automatic slice interpolation. In the last stage, 3D reconstruction of tumor volume was performed.

Distribution similarity-based approaches were applied in [PML+05]. Authors proposed two distribution similarity-based metrics for lesion detection, considering the AD problem as classification of normal/disease in 3D medical images. In the fist case, they utilised the Mahalanobis distance and the Kullback–Leibler (KL) divergence to compute the divergence between two spatial distributions of the region of interest in an image of a new subject. Each of the considered classes is represented by historical data (e.g., normal versus disease class). In the second case, maximum likelihood is adopted in order to predict the class that most likely produced the dataset of the new subject. KL divergence, as well as maximum likelihood based methods, seems to outperform the Mahalanobis distance based method. For the estimation of the probability distribution of the region of interest in 3D data space, they either estimate the mean

and covariance of the dataset or apply a semi-parametric method, namely the Expectation-Maximization method (EM) and k-means algorithm. Their method was applied to Alzheimer's disease fMRI data.

Neural Networks have also been applied to anomaly detection [DMTV03], [BFN05], [MR05].

In [DMTV03], an approach based on multiple classifier systems for classification of microcalcification was proposed. A multiple classifier system composed of two experts which are "acting" in parallel: the "μC-Expert", for the classification of the single microcalcification and the Cluster Expert for classification of the cluster considered as a whole. The final classification decision is taken based on the above decisions and applying a suitable scheme, which is based on the reliability of the evaluation of each classification. For both experts, backpropagation neural networks were utilised. In [MR05] authors proposed Radial basis function neural networks to detect anomalies in lungs. In [BFN05] a fuzzy neural network architecture was proposed for tumor detection on brain MRI images.

In many papers Support Vector Machines (SVMs) were utilised for abnormality detection [CSK+07, MLB+05, SC05, JSR+20]. For example, in [CSK+07] a Support Vector machine is utilised for detection of normal mammograms. In this paper, in order to cope with the poor separability and the overlap of the feature distributions, authors propose an uncrossed-feature and Local Probability Difference based SVM learning system to separate crossed (normal/abnormal) features. The cross distributed feature pairs were identified and mapped into new features that can be separated by a zero-hyperplane.

One-class SVM (OCSVM) is one of the most well-known one-class classification methods and was also utilised for abnormality detection in medical imaging [ZCCK05, JSR+20]. In [JSR+20] authors proposed an outlier-detection-based framework which generated the tumor masks for each image slice based on anomaly detection using independent OCSVM. The method was applied to T1-weighted (T1w) and T2-weighted-fluid-attenuation-inversion-recovery (T2-FLAIR) images to segment brain tumours.

Subspace based methods have also been applied [PATB96]. In [PATB96], authors propose a classification based model, which makes use of Principal Component Analysis (PCA), in order to detect/classify anatomically different types of linear structures in order to detect abnormal

line patterns.

Atlas-based segmentation approaches [VLMV+01], [CKPR99], [PBHG04] have also been used for abnormality detection. For instance, in [CKPR99], a 3D hierarchical deformable registration algorithm is applied to register a standard atlas to a patient's atlas. Asymmetry detection in patient's data is considered as an indicator for the existence and localisation of a pathology.

Relevance Vector Machine (RVM) [Tip01] has also been utilised for the detection of clustered microcalcifications in mammograms which can be an early sign of breast cancer. In [LYN+05], a supervised two-stage RVM network with a linear kernel is proposed as a classifier. The clustered microcalcifications (MC) is formulated as a binary classification problem and the RVM classifier determines whether an MC object is present or not.

In [XWLLP06], the abnormality detection problem in retinal images is treated as a Multiple Instance Learning (MIL) problem. The input training images are pre-processed using View-PointMiner to extract relevant features and then multiple instance learning is applied to classify an image as normal or abnormal. Also, in [QLC+16, ITW18] algorithms relying on Multiple Instance learning were proposed with application domains the abnormality detection in digital mammography and histological images respectively.

Genetic algorithms were also utilised for the detection of a pathology. In [KT07], authors detect the presence of microcalcifications in breast tissue using a genetic algorithm. After enhancement and normalisation of breast images, they detect the breast border and the nipple position using a genetic algorithm. Then they discover the suspicious regions on digital mammograms based on asymmetries between left and right breast images.

Table 2.2 lists studies that were applied in medical AD, using conventional pattern recognition approaches.

Reference	**Application domain**	**Modality**	**Method**
[BS73]	chest tumors liver tumors	X-Ray Isotope Scan	Dynamic Programming
[VMT98]	Ovarian tumours	Color Doppler Imaging	MLP
[ZLCH03]	Skull base tumor (nasopharyngeal carcinoma)	MRI	fuzzy c-means

[CKPR99]	brains with pathologies	MRI, CT	Atlas Based Deformable Registration
[DMTV03]	Breast tumours	Mammography	Multiple classifier systems
[MR05]	lung tumours	electrical impedance tomography	Radial basis function neural networks
[ZCM+99]	Breast tumours	Mammography	Discrete Wavelet Transform-based Markov Random Field (MRF)
[LYN+05]	clustered microcalcifications	mammograms	Relevance Vector Machine
[PATB96]	anatomically important linear structures	X-ray mammograms	classification model based on PCA
[PBHG04]	Brain tumours	MR	Atlas-based
[JSR+20]	Brain tumours	MRI	OCSVM
[VLMV+01]	Brain tumours	MRI	Atlas based
[XWLLP06]	Retinal Imaging	OCT	MIL
[CSK+07]	Breast tumours	Mammogram	SVM-based method
[KT07]	Breast tumours microcalcifications	Mammogram	Genetic algorithm
[MLB+05]	Brain tumor	single voxel H magnetic resonance spectra	LDA, SVM, ensemble methods
[QLC+16]	Breast cancer	Digital Mammography	SVM, MI-SVM, MILBoost
[ITW18]	Breast and colon cancer	Hematoxylin & Eosin histopathology images	attention based MIL
[FGG09]	Multiple Sclerosis lesions (brain)	MRI	GMMs and Curve Evolution
[Ger03]	Brain Tumour	MRI	MRF
[LSM+05]	Brain Tumour	MRI	CRF-SVM
[WMAG07]	Multiple sclerosis lesion detection	MRI	GMM
[WWILT+06]	Multiple sclerosis lesion detection	MRI (multichannel)	k-nearest neighbor(k-NN)
[FVPT12]	outlier detection	fMRI	Local Component Analysis

[NJW14]	Multiple sclerosis lesion detection	MRI	histogram-based
[ECKK21]	brain segmentation	MRI	KDE-DNN
[PML+05]	brain lesion	fMRI	Distribution similarity KL, Mahalanobis
[YFH+13]	liver (lesion)	CT	Naive Bayes Nearest Neighbours
[BFN05]	brain tumor	MRI	fuzzy Neural Network
[ZCCK05]	brain tumor	MRI	OCSVM

Table 2.2: A list of works in medical anomaly detection before Deep Learning

Deep Learning has been widely applied in abnormality detection in medical imaging, improving significantly anomaly detection performance. Various deep learning architectures such as CNNs, RBMs, Deep Convolutional AEs and its variants, VAEs, GANs have been utilised in order to solve the AD problem. Below, the most important works that have been presented in medical imaging literature, are discussed and sorted by the deep learning architecture.

Convolutional neural networks (CNNs) have been widely applied in deep AD, either as a feature extractor or as a classifier in a fully, weakly or semi-supervised scenario. In most of these cases, the anomaly detection problem is considered as a two-class or multiclass classification problem. In cases that CNNs are used as feature extractors, usually a pre-trained or trained from scratch state-of-the-art model such as VGG [SZ15], AlexNet [KSH12], ResNet [HZRS16] is utilised to extract low-dimensional features. These would retain the discriminative information that helps the classifier to separate anomalies from normal subjects, at the next stage. For example, in [LCW+17] multimodal pre-trained CNNs (16-layers VGG [SZ15], 50-layers ResNet [HZRS16], 22-layers GoogleNet [SLJ+15]) have been utilised, either as classifiers or as feature extractors, in order to (a) distinguish between cancerous and non-cancerous tissues and (b) to distinguish between clinically significant and indolent prostate cancer. In this work, CNNs are used as single classifiers in a supervised classification setting. Also, features which are extracted from the CNNs, are fused with hand-crafted features and they are fed as input into a SVM classifier.

In [ZXP+20] authors consider the viral pneumonia detection problem as one class classifica-

tion problem. They propose an anomaly detection framework which consists of an anomaly detection and a confidence prediction module. The anomaly detection module is composed of a pretrained on Imagenet EfficientNet [TL19], which is used as feature extractor, and a multi-layer perceptron with three 100-neuron hidden layers and a one-neuron output layer. The Confidence network consists of four 100-neuron hidden layers and a one-neuron output layer.

CNN as a classifier is utilised also in [KBM+20] for automated detection of Crohn's disease ulcers by video capsule endoscopy and in [AHAA+19] for automated ulcer detection in wireless capsule endoscopy images.

Ensembles of CNNs also have been applied to medical diagnosis [ACZ+20] [IZ18]. In [IZ18] an ensemble of three deep convolutional neural networks was used for Alzheimer's disease diagnosis (multi-class problem) from brain MRI images. The classification output of the three individual models are fused using majority voting. In another work [ACZ+20] an ensemble of deep CNNs (ResNets) is trained to identify recommended cardiac views (a view classifier) and to distinguish between normal hearts and complex congenital heart disease from screening ultrasound.

CNNs have also been applied in a weakly-supervised scenario [IGV+18]. In this work authors propose a CNN (WCNN) for weakly-supervised learning from gastrointestinal images in endoscopic video frame sequences. Images are weakly-annotated i.e. image-level instead of pixel-level labels are used. WCNN classifies images as normal or abnormal (having gastrointestinal anomalies). Furthermore, the feature maps of a deeper WCNN convolutional layer is used to detect salient points in the abnormal images.

CNNs can also be combined with other pattern recognition methods for out-of-distribution detection. For instance, in [LLDL20] a deep Isolation Forest method combined with pretrained deep CNN was applied for out-of-distribution detection in skin images. Initially, a CNN is trained for classification of normal cases. Then the hidden representation from the last convolutional layer of a pretrained CNN is utilised for constructing different Isolation Forest models for each class. The final normality score is computed as the maximum score among the isolation forests models scores.

Convolutional Autoencoders and its variants have been used either as low-dimensional feature extractors (latent space) or for deriving feature representations for generic normality

learning.

In the former category there are works in a variety of pathologies such as [SWK+19, ZWH+18, RSNL+19, AJBL20]. For instance, in [SWK+19] a deep denoising autoencoder is trained on healthy data and based on the learned feature representation, the distribution of healthy subjects is estimated, using OCSVM with a linear kernel. The method was applied on Optical Coherence Tomography to identify disease biomarkers. Similarly, in [AJBL20], for detection of subtle epilepsy lesions in multiparametric MRI, a siamese neural network, composed of stacked convolutional autoencoders has been applied to healthy scans only, in order to learn the normal brain representations. Then the middle layer representations of the above networks are fed into OCSVM models at voxel-level for classification between normal/abnormal subjects.

In the later case where autoencoders are used for normality learning, models are usually trained only utilising normal data and the basic assumption is that since they model normality, they fail to reconstruct accurately the abnormal cases during the test phase. Such works include [SHN+18, MXW+20, HYZ+20]. In [SHN+18], a 3D CAE is trained on normal cases of emergency head CT volumes. Mean squared error is considered as the abnormality score. Similarly, in [MXW+20], authors proposed an AE for abnormality detection in chest X-Ray images, trained only on normal data. Additionally, estimation of reconstruction uncertainty in each pixel is also derived. The reconstruction error normalised by uncertainty is utilised as abnormality score. In [HYZ+20] authors proposed an encoder-decoder-encoder architecture for simultaneously optimizing entropy and mutual information. The encoder-decoder (input-reconstruction of input) is focusing on optimizing the mutual information, while the second encoder (reconstruction of input-latent space) is focusing on optimizing the entropy. The two encoders are enforced to share similar encoding with a consistent constraint on their latent representations. The anomaly score is computed based on the reconstruction error and the entropy. The method is applied both for detection of metastases in digital pathology and recognition of chest diseases on the chest X-rays in the NIH dataset [WPL+17].

In [BWAN20] bayesian-skip autoencoder with Monte-Carlo dropout (to estimate epistemic uncertainty), has also been applied to brain MRI lesion unsupervised detection. The residual error (i.e $l1$ norm between input and mean of n MC reconstructions) is used as anomaly score.

Finally, in [BWAN19] a spatial autoencoder together with a segmentation network were utilised

for white matter lesion unsupervised segmentation.

Generative Models such as *Variational Autoencoders (VAE)* and *Generative Adversarial Networks (GAN)*, have also been used in AD.

Some of the most important works which utilise *Variational autoencoders* [KW14] for anomaly detection include [LX18, TTHX19, ZLS+19, CK18, BWAN20, CYTK20, CP+19, ZDW20]. In [BWAN18, BDW+20] authors proposed a combination of a spatial variational autoencoder with a discriminator, trained in a GAN-likewise approach. It is applied to high resolution MRI images for unsupervised lesion segmentation. AnoVAEGAN uses a variational autoencoder and tries to model the normal data distribution that will lead the model to fully reconstruct the healthy data, while it is expected to fail to reconstruct abnormal samples. The discriminator classifies the inputs as either real or fake (reconstructed data). The aim of the discriminator is to improve the reconstructed samples realism. As anomaly score the $l1$ norm of the difference between the original image and the reconstructed image is used. VAE is also applied in [LX18], trained only on normal data in order to detect melanoma on skin images. Similarly, in [CK18], a VAE (and an Adversarial Autoencoder [MSJ+16]) along with a constraint to encourage latent space consistency during training is applied in order to detect lesions in brain MRI. Finally, in [ZDW20] VAE and β-VAE have been proposed for brain MRI lesion detection.

Generative Adversarial Networks (GANs) [GPAM+14] have been widely applied to deep anomaly detection in medical imaging [GZZ+20, SCWZ20, ZGC+20, SSW+19, ASF+19, FKM+20, BGW+20, HRM+20, SZT+19, AMSCK17, TYBHX19, TTHX19, WCXM20]. GANs for anomaly detection were first proposed by [SSW+17]. In [SSW+17], a deep convolutional generative adversarial network *AnoGAN*, inspired by DCGAN [RMC16], is used. During the training phase, only healthy samples are used. This approach consists of two models. A generator, which generates an image from random noise and a discriminator, which classifies real or fake samples as in common GANs. In their work, a residual loss is introduced, which is defined as the $l1$ norm between the real images and the generated image. This enforces the visual similarity between the initial image and the generated one. Furthermore, in order to cope with GAN instability, instead of optimizing the generator parameters via maximizing the discriminator's output on generated examples, the generator is forced to generate data whose intermediate feature representation of the discriminator is similar to those of the real images. This is defined as the $l1$

norm between intermediate feature representations of the discriminator given as input the real image and the generate image respectively. In AnoGAN, an anomaly score is defined as the loss function at the last iteration, i.e the residual error plus the discrimination error. AnoGAN has been tested on a high-resolution SD-OCT dataset.

Similar to AnoGAN, a faster approach, f-AnoGAN has been proposed in [SSW+19]. In this work, the authors train a GAN on normal images, however instead of the DCGAN model a Wasserstein GAN (WGAN) [ACB17] [GAA+17] has been used. Initially, a WGAN is trained in order to learn a non-linear mapping from latent space to the image space domain. Generator and discriminator are optimised simultaneously. Samples that follow the data distribution are generated through the generator, given input noise sampled from the latent space. Then an encoder (convolutional autoencoder) is trained to learn a mapping from the image space to the latent space. For the training of the encoder, different approaches are followed. Both [SSW+17] as well as [SSW+19] use image patches for training and are modular methods which are not trained in an end-to-end fashion.

Another GAN-based method applied to OCT data has been proposed by [ZGC+20], in which authors propose a Sparsity-constrained Generative Adversarial Network (Sparse-GAN), a network based on an Image-to-Image GAN [IZZE17]. Sparse-GAN consists of a generator, following the same approach as in [IZZE17], and a discriminator. Features in the latent space are constrained using a Sparsity Regularizer Net. The model is optimized with a reconstruction loss combined with an adversarial loss and sparsity regularization. The anomaly score is computed in the latent space, not in the image space. Furthermore, an Anomaly Activation Map (AAM) is proposed to visualise lesions.

Encoder-decoder-encoder framework, trained adversarially together with a discriminator is applied in [TTHX19, TYBHX19]. The adversarial one-class model was proposed for chest radiograph one-class anomaly detection. In [TYBHX19] the encoder-decoder model is composed of a U-Net like autoencoder while in [TTHX19] by a 5-layer CNN. In [TTHX19], anomaly score is computed based on the chi-square (χ^2) distance between two latent vectors, and the difference between the real input image and its generated fake image. However, anomaly score in [TYBHX19] is computed based on three terms, namely the distance between two latent vectors, the difference of the real input image and its generated fake image and the output of

discriminator (likelihood that a generated fake image does not look realistic).

In [SCWZ20] adGAN, an alternative framework based on GANs, is proposed. The authors introduce two key components: fake pool generation and concentration loss. adGAN follows the structure of WGAN and consists of a generator and a discriminator. The WGAN is first trained with gradient penalty using healthy images only and after a number of iterations, a pool of fake images is collected from the current generator. Then a discriminator is retrained using the initial set of healthy data as well as the generated images in the fake pool, with a concentration loss function. Concentration loss is a combination of the traditional WGAN loss function with a concentration term which aims to decrease the within-class distance of normal data. The output of the discriminator is considered as anomaly score. The method is applied to skin lesion detection and brain lesion detection.

In another work [GZZ+20], a combination of WGAN and CNN network and a variant of ALOCC [SKFA18] one class classification algorithm was utilised for fetal congenital heart disease detection. During training, both normal and disease samples were used. Adversarial one-class classification method (which is utilised for screening end-systolic (four chamber heart) video slices) was combined with video transfer learning in order to improve the performance of the detection system.

Adversarially learning similar to ALOCC work [SKFA18] is also applied in [ASF+19] for detecting irregular tissues in mammography images. Two networks were adversarially trained and an irregularity score function was defined during the test phase in order to detect abnormal images.

GANs were also proposed in [FKM+20] and [AMSCK17]. They were applied to breast ultrasound imaging and brain MRI lesion detection, respectively. The anomaly score was based on both the reconstruction error and discriminator output in the first case while solely on discriminator output in the second case.

Image-to-Image translation with GANs has also been applied for detection and localisation of abnormalities in medical images. In [BGW+20] a Cycle-GAN based model was proposed for unsupervised anomaly segmentation in brain MRI images. The model is trained to translate healthy brain images between distributions of only healthy data in different styles. The anomaly

score is defined as the residual between input and its reconstruction (after a complete cycle).

In [SZT+19] a Fixed-point GAN model was proposed for disease detection and localization. The model is trained for the cross-domain translation (diseased images to healthy images and vice versa) as well as same-domain translation. Image-level annotations are utilised for training Fixed-point GAN. The maximum value in the difference map (subtracting the translated healthy image from the input image) across all the pixels is defined as the detection score.

Finally, in [HRM+20] a two-step method was proposed for the detection of brain anomalies at different stages on multi-sequence structural MRI using an image-to-image GAN-based multiple adjacent brain MRI slice reconstruction.

Other generative models such as deep belief networks have also been used. For instance, in [YTB+17] a four-layer deep belief network (DBN) [HOT06] is applied to 3D image patches of T1-weighted (T1w) MRIs and myelin maps to learn latent joint feature representations. The extracted features are then used to train a random forest in order to discriminate images of subjects suffering from multiple sclerosis from normal subjects. In Table 2.3 medical AD studies using deep learning-based methods have been listed.

In Figure 2.3 a taxonomy of methods which have been applied in the medical imaging domain is presented.

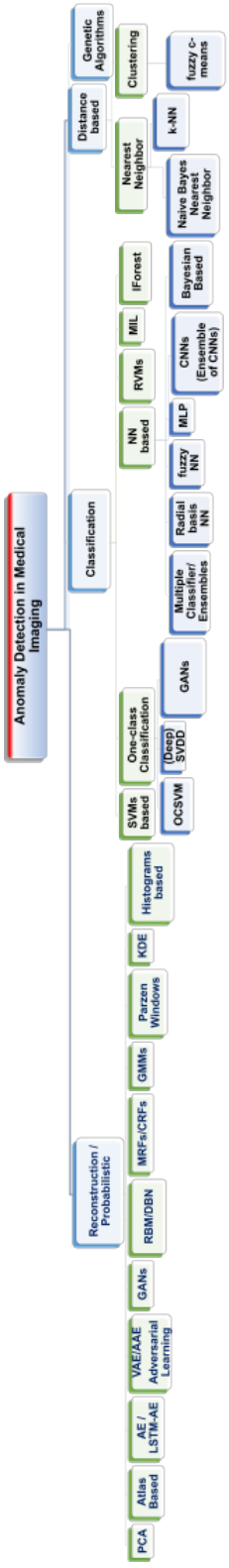


Figure 2.3: A taxonomy of anomaly detection methods in medical imaging

Reference	Application dom	Modality	Availability of Labels	Method
[LX18]	Skin	Dermoscopy data	Unsupervised	VAE
[ACZ+20]	Fetal Heart	Ultrasound	Supervised	Ensemble of CNNs
[BWAN18]	Brain	MR	Unsupervised	spatial VAE-discriminator (AnoVAEGAN)
[SWK+19]	Retinal	OCT	Unsupervised	Deep DAE-OCSVM
[TYBHX19]	Chest	X-Ray	One-class classification	Encoder-Decoder-Encoder & Discriminator Adversarial Learning
[GZZ+20]	Fetal Heart	Echocardiography	One-class classification	GAN-based
[SCWZ20]	Skin, Brain	MRI/Dermoscopy data CT-scan	Unsupervised	GAN-based Adversarial Learning
[BWAN19]	Brain	MR	Supervised / Unsupervised	Spatial AE
[SHN+18]	Head	CT	Unsupervised	3D CAE
[LLDL20]	Skin	Dermoscopy data	Out-of-Distribution Detection	pretrained CNN Deep isolation Forest
[CK18]	Brain	MRI	Unsupervised	Constrained Adversarial AE
[ZGC+20]	Retinal	OCT	Unsupervised	Sparsity-constrained GAN
[IGV+18]	Gastrointestinal	Gastrointestinal Endoscopy	weakly-supervised	CNN
[SSW+17]	Retina	SD-OCT	Unsupervised	GAN-based
[SSW+19]	Retina	OCT	Unsupervised	GAN-based
[ASF+19]	Breast	Mammography	Unsupervised	Adversarial learning
[FKM+20]	Breast	Ultrasound	Unsupervised	GAN
[YTB+17]	Brain	T1-w MRI myelin water imaging	Unsupervised	DBN Random Forest
[USHE19]	Brain, Lungs	MRI, CT	Unsupervised	Conditional VAE
[ZKP+18]	Brain	MRI	Unsupervised	Context-encoding VAE
[LCW+17]	prostate	MRI	Supervised	CNN-SVM
[IZ18]	Brain	MRI	Supervised	Ensemble of CNNs
[HWL+18]	Neural foraminal	MRI	supervised	Mutli-task CNN
[KBM+20]	Bowel ulcers	capsule endoscopy	Supervised	CNN

[AHAA+19]	Ulcer	wireless capsule endoscopy	Supervised	CNN
[SKS+19]	renal	DW-MRI	Supervised	CNN-AE
[ZWH+18]	Brain	functional-MRI	Supervised	AE-SVM
[BMVT20]	Chest	X-Rays	Self-Supervised	Aggregation based U-Net
[BWAN20]	Brain	MRI	Unsupervised	Bayesian skip-AE
[BGW+20]	Brain	MRI	Unsupervised	CycleGAN
[AJBL20]	Brain	MRI	Unsupervised	siamese CAE-OCSVM
[RSNL+19]	prostate	PET/CT	Unsupervised	CAE Density estimation
[CYTK20]	Brain	MRI	Unsupervised	VAE/GMVAE
[MXW+20]	Chest	X-Ray	Unsupervised	AE
[ZXY+20]	Retinal	OCT	semi-supervised	CNN based
[CP+19]	Brain	MRI	Unsupervised	VAE
[WSC20]	chest	X-Ray	weakly supervised	GAN-based
[ZDW20]	Brain	MRI	Unsupervised	VAE, β-VAE
[ZLS+19]	Breast	mammograms	Supervised	Convolutional Encoder-Decoder, ResNet50
[ZXP+20]	Chest	X-Ray	One-class classification	CNN-based
[HRM+20]	Brain	MRI	Unsupervised	GAN-based
[TMP+20]	Colonoscopy	Colonoscopy	Few-shot	CNN based
[SOS+20]	Retinal	OCT	weakly supervised	Bayesian-Unet
[SZT+19]	Brain/ Pulmonary Embolism	MRI/ computed tomography pulmonary angiography	weakly-supervised	Fixed-Point GAN
[HYZ+20]	Chest/ Metastases Detection	X-Ray/ Digital Pathology	Semi-Supervised	Encoder-Decoder & Mutual Information Entropy
[AMSCK17]	Brain	MRI	Semi-supervised	GAN
[TTHX19]	Chest	X-Ray	One-class learning	Adversarial learning
[TBFD20]	Chest/ Metastases Detection	X-Ray/ Digital Pathology	Semi-Supervised	perceptual AE
[OYU+19]	Retinal	OCT	Unsupervised	Isolation Forest transfer learning

[WCXM20]	Brain Tumor	MRI	Semi-supervised	Latent Regularized Adversarial Network

Table 2.3: A list of works in anomaly detection in medical imaging using Deep Learning

Datasets and Evaluation

Datasets: Both public as well as private datasets have been used for validation of anomaly detection algorithms in medical imaging.

One of the most commonly used public datasets for anomaly detection in brain is the Multimodal Brain Tumor Image Segmentation (BRATS challenge) (2017, 2019) dataset [MJB+15, BRJ+18] for lesion detection/segmentation of brain tumors. BRATS scans contain (a) native T1, (b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes, and were acquired using different clinical protocols and various scanners from multiple ($n = 19$) institutions. All the datasets have been manually segmented by at least one (maximum four) rater who follow the same annotation protocol.

Another popular dataset in brain research is the 2015 Longitudinal MS lesion segmentation challenge (organised in conjunction with ISBI 2015) which contains different subjects with T1-w(magnetization prepared rapid gradient echo–MPRAGE), T2-w, PD-w (double spin echo–DSE) and T2-w FLAIR images. All data has been acquired with 3.0 Tesla MRI scanner [CRJ+17]. The training data consisted of five subjects with a mean of 4.4 time-points, and test data of fourteen subjects with a mean of 4.4 time-points. The dataset also included manual delineations, of the white matter lesions associated with multiple sclerosis, made by two human experts. [CRJ+17]

For chest X-Ray anomaly detection the two most common datasets are the NIH X-Rays [WPL+17] and the CheXpert [Iea19] dataset.

The NIH X-Rays [WPL+17], called "ChestX-Rays8", consists of $108,948$ frontal-view X-ray images of $32,717$ unique patients with the text-mined eight disease image labels. The $84,312$ images are normal and the the rest present one or more pathologies. Each image can have multiple labels.

body part	Dataset
brain	BRATS [MJB+15], MS lesion [CRJ+17], MOOD [ZPK+21]
breast	CAMELYON16 [EBVJvD+17]
skin	ISIC2018 Challenge [CRT+18, TRK18]
chest	CheXpert [Iea19], NIH Chest XRay [WPL+17]
abdominal	MOOD [ZPK+21]
Other	Neuropathology [NL20, FXH+18], MURA [RIB+17], LIDC-IDRI [AIMB+11]

Table 2.4: Existing benchmark datasets for medical imaging anomaly detection

The "CheXpert" dataset [Iea19] contains 224,316 entries of chest X-rays of a total of 65,240 patients. From these X-rays, they extract the observations of 14 different diseases using Natural Language Processing. Labels are positive (presence of pathology), negative (absence of pathology) and uncertain. The CheXpert validation and test sets were labeled manually by expert radiologists.

Furthermore, for skin image anomaly detection, ISIC2018 Challenge Disease Classification [CRT+18, TRK18] is commonly used. The dataset is split into three different image analysis tasks, namely: Lesion Segmentation, Lesion Attribute Detection and Lesion Disease Classification. For each task, training datasets contain labels/annotations and consist of different numbers of available images.

Additionally, the MURA dataset [RIB+17] contains upper limb X-rays images labeled regarding whether they contain anomaly or not. This dataset is composed of seven classes of body parts: finger, hand, wrist, forearm, elbow, humerus, and shoulder. There are a total of 40,561 multi-view radiographic images from 12,173 patients. Data was manually labelled as normal/abnormal by board-certified radiologists.

Finally, for detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections, CAMELYON16 dataset was proposed [EBVJvD+17]. A training data set of whole-slide images from 2 centers with ($n = 110$) and without ($n = 160$) nodal metastases verified by immunohistochemical staining were provided. Evaluation of performance of proposed algorithms is done in an independent test set of 129 whole-slide images (49 with and 80 without metastases).

Table 2.4 summarises the existing benchmark datasets for anomaly detection in medical imaging.

Evaluation: For quantitative evaluation of anomaly detection algorithms, performance metrics usually are presented. The most common presented metrics are:

- *True Positive Rate (TPR) or Recall*: is the proportion of positive classes from the total possible positive conditions that are True Positives(TP) and False Negatives (FN). Precisely, TPR=$\frac{TP}{TP+FN}$. Informally, in an anomaly detection context, Recall is the fraction of all real anomalies that are successfully detected [TLZ+18].

- *Precision (PR)*: PR=$\frac{TP}{TP+FP}$. Precision is informally the fraction of all detected anomalies that are real anomalies [TLZ+18].

- *F1 Score*: Precision and Recall are complementary and can be combined. F1 score is defined as the harmonic mean of Precision and Recall (TPR). F1 score=$\frac{2*PR*TPR}{PR+TPR}$.

- *True Negative Rate (TNR) or Specificity*: is the proportion of correctly identified negative classes from the total possible negative conditions, that are true negative (TN) and false positive (FP). TNR=$\frac{TN}{TN+FP}$

- *False Positive Rate (FPR)*: FPR is the rate of False alarms, i.e mis-classifying some normals as outliers FPR= $\frac{FP}{FP+TN}$ =1-TNR, where FP the False positives.

- *Receiver Operating Curve (ROC) and Area Under Curve (AUC)*: Receiver Operating Curve (ROC) is commonly used to measure the performance of the classifier by plotting true positive rate against false positive rate. The area under this curve, AUC, is a measure of the quality of the detector. A higher value of AUC of ROC shows that the model is performing well. An AUC value of 1 indicates the best performance, while a value of 0.5 indicates a random prediction. This metric is threshold-invariant as well as scale-invariant.

For qualitative analysis of anomaly detection algorithms, usually activation or localisation maps are presented. Visualisation maps are derived either implicitly by the model and its attention mechanisms [SOS+19, ZGMO19] [ZGC+20] or using gradient- and optimisation- based methods [SCD+17].

2.2 Uncertainty Estimation in Deep Neural Networks

2.2.1 Introduction

In many safety-critical applications, a decision which will be derived by a model, such as a deep learning model, is crucial and could have potential negative impact on human beings e.g self-driving cars, medical diagnostics. For this reason, a key question is, how confident the model is in its decision and consequently to which extent we could trust the result of an algorithm. In general, uncertainty is a sign of abstention from the prediction which could also play an important role in anomalies detection. Estimation of uncertainty in a model output can originate either from the data (aleatoric uncertainty) or the model itself (epistemic uncertainty). Although deep learning methods show an outstanding performance, there are limitations in the application of these methods to computer vision problems, such as lack of interpretability, they are prone to overfitting, easily fooled by adversarial examples, poor at representing uncertainty. For instance, adding perturbation on images, may lead to misclassification of the natural images [MDFFF17]. Thus, it is very important to be able to give a measure of confidence for the result along with the model output. *Uncertainty* can be categorised in: *Aleatoric* uncertainty and *Epistemic* uncertainty [Gal16]. Both types of uncertainties can form the predictive uncertainty.

Aleatoric Uncertainty captures noise inherent in the observation [KG17, CMK+20], e.g. sensor noise, label disagreement. It refers to the randomness of the data generating process and cannot be explained away with more data. Furthermore, uncertainty does not increase for out-of-data examples [KG17]. Aleatoric uncertainty is very important to be modeled for large datasets and for real-time applications [KG17]. Aleatoric uncertainty can be further categorised into *Homoscedastic uncertainty* and *Heteroscedastic uncertainty*. In the former sub-category uncertainty remains constant for different inputs, while in the latter one, some inputs have potentially more noisy outputs than others, i.e it is data-dependent uncertainty from observation noise.

Epistemic or Model Uncertainty is caused by ignorance about the model that generated the data, including uncertainty in the model, parameters and convergence. Epistemic uncertainty is reducible if more information i.e data/measurements from the low density regions, become available. In any case, it cannot be removed entirely. Epistemic uncertainty increases with

decreasing training size and increases with examples which lie far from the training data distribution. As a result, it is very important to model epistemic uncertainty in cases where available datasets are small and contain limited training data and also in safety-critical applications, in order to understand and better interpret the model's outcome.

Distributional Uncertainty is also known as dataset shift [QCSSL09]. It refers to the uncertainty which will occur from mismatch in data distributions between the training dataset and the testing dataset (distributional mismatch) [MG18].

One of the first attempts to estimate reliability and confidence of Artificial Neural Networks was made in [LKLHU92], where an extension of radial basis function networks (RBFNs), called "validity index network" was introduced. It calculates the reliability and confidence by implementing additional output nodes in the RBF network [BL88].

Most existing common methods to estimate uncertainty in deep learning models are based on Deep Bayesian Neural Networks (BNNs) [TLS89, Mac92], Deep Ensembles [LPB17] and Monte-Carlo Dropout (MC-Dropout) [MRR$^+$53, GG16b].

2.2.2 Probabilistic Modeling in Deep Neural Networks

Introduction to Bayesian Inference

Let us assume a fully annotated dataset of images and annotations $D := \{\mathbf{x}_n, \mathbf{y}_n\}$ with $\mathbf{x}_n \in \mathbb{R}^p$ and $\mathbf{y}_n \in \mathbb{R}^q$ for all subjects $n = 1, 2, ..., N$. We further assume that N samples in the dataset are independent and identically distributed (i.i.d). The target vector $\mathbf{y}$ can be either real valued (regression) or categorical (classification). Dataset D can be split into D_{train} and D_{test} for the training and testing phases respectively. We make the assumption of a parametric distribution (probabilistic model) of the form $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ and estimate the parameters $\boldsymbol{\theta}$ which maximise the distribution. Using maximum likelihood estimation (MLE) we can find the parameters $\boldsymbol{\theta}$ which maximise the likelihood function, i.e

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}) \overset{\text{i.i.d}}{=} \arg\max_{\boldsymbol{\theta}} \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\theta}) \tag{2.1}$$

Alternatively, adding prior information about parameters, $p(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ is estimated using maximum a posteriori estimation (MAP):

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D) = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2.2}$$

In cases where the model is a neural network, the parametric distribution $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is defined as $p(\mathbf{y}, f(\mathbf{x}; \boldsymbol{\theta}))$, where $f(\mathbf{x}; \boldsymbol{\theta})$ is a neural network. However, both MLE and MAP approaches give a point estimate of the parameters $\boldsymbol{\theta}$. Consequently, the output of a NN is just a single value (deterministic output) without any information about the confidence of the model's output. Quantifying uncertainty using Bayesian inference is widely applied in deep learning models. We initially introduce the key concepts of Bayesian inference and then Bayesian Neural Networks (description and uncertainty estimation) will be discussed.

The two basic elements for estimating the posterior distribution based on Bayes theorem is the prior $p(\boldsymbol{\theta})$ and the likelihood function $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. The former expresses the belief about parameters values $\boldsymbol{\theta}$ (prior distribution $p(\boldsymbol{\theta})$) before observing the data. Based on Bayes theorem, the posterior distribution of parameters $\boldsymbol{\theta}$ is defined using the product of prior and likelihood:

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{x})} \tag{2.3}$$

The denominator, *marginal likelihood or model evidence*, is used to ensure that the posterior is normalised.

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.4}$$

For the estimation of the predictive distribution, given a new point $\mathbf{x}_*$, we compute:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})d\boldsymbol{\theta} \tag{2.5}$$

Marginalisation can be computed analytically, in case the likelihood is *conjugate* to the prior distribution (in the the marginal likelihood) [Gal16]. However, in many cases, computing posterior distributions and consequently predictive distributions (Eqs. 2.4 and 2.5), is often analytically intractable (i.e have no closed form solution) and thus an approximate inference method is required.

There are several approximation techniques such as Laplace approximation [Bis06, DL90], stochastic approximation methods (which are based on numerical sampling methods-Monte Carlo methods) and deterministic approximations such as Variational inference methods.

Laplace approximation [Bis06, DL90] finds a mode of the posterior distribution and then an approximation is constructed with a normal distribution by the second order Taylor expansion about the mode. More precisely, Laplace approximation is based on computing the mode of the posterior $\hat{\boldsymbol{\theta}}_{MAP}$ and sets a posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mu, \Sigma)$ with $\mu = \theta_{MAP}$. Σ is computed based on the negative inverse Hessian evaluated at the MAP solution, i.e

$$\Sigma = -\Bigg[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} logp(\boldsymbol{\theta}|D)|_{\boldsymbol{\theta}=\theta_{MAP}}\Bigg]^{-1}$$

After computation of θ_{MAP} and Σ, we can either take Monte Carlo samples from the approximate posterior (Sampled Laplace) or linearise the Gaussian model [FLHLT19].

Sampling methods: Numerical sampling (Monte Carlo methods) is a way to cope with intractable integrals in Bayesian learning. We can draw samples from the posterior distribution, and using the empirical distribution to estimate all the appropriate quantities. Some of the most well-known methods include Rejection sampling [Bis06], Importance Sampling [Bis06] and Markov Chain Monte Carlo methods(MCMC) [MRR$^+$53].

Both Importance and Rejection sampling are not efficient in high dimensional spaces. MCMC methods are based on the construction of a Markov chain to step through a high dimensional posterior probability distribution, $q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$ as the proposal distribution. It is a stationary distribution which converges to the desired posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

The most widely applied MCMC methods for Bayesian Inference include the Metropolis-Hastings algorithm [MRR$^+$53] [Has70], Gibbs sampling [Bis06] and the Hamiltonian Monte Carlo [Nea96].

Metropolis-Hastings algorithm [MRR$^+$53] [Has70] is a random walk and an acceptance rule (i.e Metropolis-Hastings ratio) is utilised to converge to target distribution. Beginning from a starting point θ_0 from a distribution $p_0(\boldsymbol{\theta})$, we sample a $\boldsymbol{\theta}^*$ from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{k-1})$ for $k = 1, 2, ...$ Then, based on the Metropolis-Hastings ratio, the new θ, $\boldsymbol{\theta}^k$ is either accepted ($\boldsymbol{\theta}^*$) as the next sample or rejected ($\boldsymbol{\theta}^{k-1}$-retaining the previous state).

A special case of the above algorithm is the Gibbs sampling [Bis06] method. In Gibbs sampling the parameter θ is divided into d components i.e $\boldsymbol{\theta} = \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_d$. At each iteration $\boldsymbol{\theta}_j^k \sim p(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j}^{k-1}, y)$. So $\boldsymbol{\theta}_j$ is sampled based on the latest values of each component, as it is described above. The main concept of Gibbs sampling is to generate posterior samples by cycling through each variable to sample from its conditional distribution with the remaining variables fixed. Usually samples can be correlated as well as the convergence is slow. Hamiltonian Monte Carlo(HMC) is another method which is based on Hamiltonian dynamics [Nea96] and is a more efficient approach.

Often, all the above methods are not scalable to large scale datasets or to large deep neural networks. In order to solve this issue, Stochastic Gradient MCMC (SG-MCMC) [MCF15] has been proposed as well as many variants of it such as Stochastic Gradient Langevin Dynamics (SGLD) [WT11], Stochastic Gradient HMC [CFG14] and cyclic stochastic gradient MCMC [ZLZ+20].

Variational Inference: Using Variational inference we approximate the posterior $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ by a simpler distribution $q(\boldsymbol{\theta})$ (variational distribution). Through variational inference, the approximation problem is transformed to an optimization problem. This can be formulated as:

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in Q} f(q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})) \tag{2.6}$$

where Q is the set of all closest distributions to p. The goal is to find the "best" distribution from a set of "least" far distributions as shown in Figure 2.4. f measures the discrepancy between the two distributions q and p. Discrepancy is "translated" in divergence between distributions.

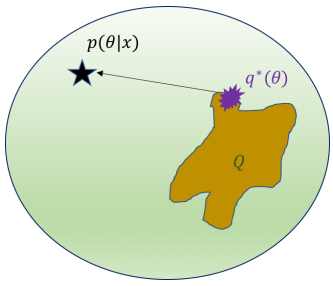


Figure 2.4: Variational Bayes

The most common similarity measure is the *Kullback-Leibler* (KL) divergence [Kul59, KA51].

KL divergence of two distributions is defined as:

$$KL[q(\boldsymbol{\theta})||p(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta}) log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} = \mathbb{E}_q\big[log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\big] \tag{2.7}$$

Based on the above equation, the KL divergence is low in case q is low (regardless of p) or in case q is high and p is also high. In the case that q is high and p is low then the KL divergence is high. One of the most important properties of KL divergence is that it is not symmetric, i.e $KL[q||p] \neq KL[p||q]$. The lower the KL the more similar the two distributions are. KL is always non-negative, $KL \geq 0$, and it is equal to zero if and only if $p = q$. Fitting q to p (reverse KL-exclusive KL) using KL will give different behaviour than minimizing p to q (forward KL-inclusive KL).

Variational Bayes [HvC93, BCKW15, KW14, Cha18] is a technique of Variational Inference which minimizes the KL divergence between a variational distribution q_ϕ with ϕ variational parameters and the posterior distribution of the parameters $\boldsymbol{\theta}$ given a dataset D, $p(\boldsymbol{\theta}|D)$. More precisely the $KL[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|D)]$ is minimised and using logarithmic identities and the Bayes theorem ($P(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}$) we can derive the following computations:

$$KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta}|D)] = \int q_\phi(\boldsymbol{\theta}) log \frac{q_\phi(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|D)} d\theta \tag{2.8}$$

$$= \int q_\phi(\boldsymbol{\theta}) log \frac{q_\phi(\boldsymbol{\theta}) p(D)}{p(D|\boldsymbol{\theta}) p(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{2.9}$$

$$= \underbrace{KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta})] - \mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[logp(D|\boldsymbol{\theta})\big]}_{\text{-ELBO}} + \underbrace{logp(D)}_{\text{const. in q}} \tag{2.10}$$

Rewriting the above Equation:

$$logp(D) = KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta}|D)] - KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta})] + \mathbb{E}_{q_\phi(\boldsymbol{\theta})}[logp(D|\theta)] \tag{2.11}$$

$logp(D)$ can be also written as:

$$logp(D) = log \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.12}$$

$$= log \int \frac{q_\phi(\boldsymbol{\theta})}{q_\phi(\boldsymbol{\theta})} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.13}$$

Applying Jensen's inequality we have:

$$logp(D) \geq \int q_\phi(\boldsymbol{\theta}) log \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q_\phi(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{2.14}$$

$$= -\int q_\phi(\boldsymbol{\theta}) log \frac{q_\phi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} + \int q_\phi(\boldsymbol{\theta}) logp(D|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.15}$$

$$= -KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta})] + \mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[logp(D|\boldsymbol{\theta})\big] \tag{2.16}$$

Thus, we have:

$$logp(D) \geq \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[logp(D|\boldsymbol{\theta})\big] - KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta})]}_{\text{Evidence Lower Bound (ELBO)}} \tag{2.17}$$

ELBO consists of a data-fit term which is the expected log-likelihood $\mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[logp(D|\boldsymbol{\theta})\big]$ and a regularizer $KL[q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta})]$which shows that $q_\phi(\boldsymbol{\theta})$ should not differ from the prior $p(\boldsymbol{\theta})$. Thus, instead of minimising the KL divergence between q distribution and the posterior of parameters $\boldsymbol{\theta}$, we define a lower bound of the model evidence and maximize it.

$$q^*(\theta) = \arg\max_{q \in Q} ELBO(q) \tag{2.18}$$

A common approach in variational inference is the Mean-Field Variational Inference (MFVI) approach. In order to define the set of distributions Q, a basic assumption is made (a nonparametric restriction). This assumption is that Q (variational family) factorises $q(.) = \Pi_i^M q_i(\boldsymbol{\theta} i)$ for some partition $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M\}$ of θ. Usually, terms $q_i(.)$ are modelled as distributions from exponential family due to the conjugacy, however any other distribution can also be utilised. The optimization problem for finding the optimal q^* can be solved using different methods such as Coordinate ascent variational inference [Bis06], Stochastic variational inference [HBWP13], automatic differentiation variational inference [KTR$^+$17].

Sampling vs Variational Inference: Both methods can be utilised for Bayesian Inference. However, both methods present some limitations. Monte-Carlo methods transform the approximation inference problem to a sampling problem, while variational inference to an optimization one. Sampling methods have higher computational cost. This is because of the random walk process, since for each sample we need to re-estimate the prior and likelihood in order to compute the posterior [Dep19]. Furthermore, in sampling methods there are more hyperparameters which are data and model dependent, such as the proposal distribution and the number of samples. Sampling methods can fit better in small datasets or in cases where the heavy computational cost is acceptable and are more unbiased compared to Variational inference. On the other hand, Variational Inference is a faster process, can easily be applied to large datasets and can prove easily adaptable compared to sampling methods [BKM17b].

2.2.3 Bayesian Neural Networks

Bayesian Neural Networks (BNNs) [Nea96, Mac92] place a prior distribution on the weights of the network. BNNs could provide uncertainty about the functional mean through posterior predictive distribution [YPGDV19]. Let $p(\mathbf{w})$ a prior over the (weights) parameters. Then applying Bayes theorem, the posterior distribution becomes:

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{\int_{\mathbf{w}'} p(D|\mathbf{w}')p(\mathbf{w}')d\mathbf{w}'} \propto p(D|\mathbf{w})p(\mathbf{w}) \tag{2.19}$$

Since the denominator is intractable, as discussed above, approximation methods for computing the posterior are required. The most common approaches for Bayesian deep learning inference are the MCMC methods and variational inference methods including MC-Dropout [GG16b] and Bayes-by-Backprop method [BCKW15].

From the former category, Stochastic Gradient MCMC [WT11] methods provide a promising direction for bayesian deep neural networks inference. However, since these methods imply many challenges, such as computational time, many attempts have been made in order to improve the sampling efficiency including Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [CFG14], preconditioned stochastic gradient Langevin dynamics (pSGLD) [LCCC16] and cyclical stochastic gradient MCMC [ZLZ$^+$20].

A review of existing Variational inference methods for BNNs is given in [SRV+20] and [APH+21].

Dropout was initially introduced by [HSK+12] as a form of regularization in neural networks. However, dropout can also be explained as an alternative way to perform Variational inference, and can be used at test phase as a form of ensemble learning (MC-dropout) [GG16b].

An independent random variable which follows Bernoulli distribution is introduced. For example, in a single MLP with L layers and dimension of each layer $K_j \times K_{j-1}$, the weights (W) are drawn from:

$$W_\rho = \text{diag}(\rho)W$$

$$\rho_{i,j} \sim Bernoulli(p_i) \text{ for } i = 1, ...L, j = 1, ..., K_{i-1} \tag{2.20}$$

where ρ is sampled from a Bernoulli distribution. This corresponds to randomly setting some units in the network to zero. A model is trained with dropout and during the test phase, we can sample the posterior distribution over the weights using dropout. In [KSW15] Gaussian dropout is also proposed.

Bayes-by-backprop [BCKW15] is another promising method. The proposed method minimises the Variational Free Energy. It utilises the reparametrisation trick to find an unbiased estimate of gradients of the cost function in order to learn distributions over the weights of a BNN.

2.2.4 Uncertainty Estimation in Deep Neural Networks

The predictive distribution in BNNs is defined as:

$$p(\mathbf{y}_*|\mathbf{x}_*, D) = \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|D)d\mathbf{w}$$

As discussed above, in order to alleviate the issue with the intractable true posterior $p(w|D)$, a variational distribution, $q_\phi(w)$, is utilised to approximate it. Thus,

$$p(\mathbf{y}_*|\mathbf{x}_*, D) \approx \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w})q(\mathbf{w})d\mathbf{w} \tag{2.21}$$

Instead of computing the integral over the weight space, MC sampling is applied and the estimator becomes:

$$p(\mathbf{y}_*|\mathbf{x}_*, D) \approx \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w}_t) \tag{2.22}$$

where $\mathbf{w}_t$ is a sample from the variational distribution, i.e $\mathbf{w}_t \sim q_\phi(\mathbf{w})$ and T is the number of samples. Entropy and moment based predictive uncertainty can be computed [KW+20, Gal16].

An important aspect in uncertainty estimation is to disentangle uncertainty into aleatoric and epistemic uncertainty. This is important since the source of each type of uncertainty is different (i.e data vs model). For instance, high aleatoric uncertainty could be an indicator for noisy data, while high epistemic uncertainty is a sign of data that are out-of-training distribution.

In [KG17] authors proposed a novel way to estimate epistemic and aleatoric uncertainty. The proposed model predicts the variance of the output, in addition to the output. More precisely, they constructed a BNN (placing priors over the weights) with the last layer (before activation) consisting of mean and variance of logits. Then for weights $\hat{\mathbf{w}}_{t=1}^{T}$ and the corresponding outputs $(\hat{\mu}_t, \hat{\sigma}_t^2)$, the estimators(aleatoric and epistemic) of uncertainty is given by:

$$\frac{1}{T}\sum_{t=1}^{T} \text{diag}(\hat{\sigma}_t^2) + \frac{1}{T}\sum_{t=1}^{T} (\hat{\mu}_t - \bar{\mu})^{\otimes 2} \tag{2.23}$$

where $\bar{\mu} = \sum_{t=1}^{T} \hat{\mu}_t / T$ and T the number of samples [KW+20].

Predictive uncertainty also can be decomposed into epistemic and aleatoric uncertainty following [DHLDVU18, SG18, HHGL11, Gal16]:

$$\underbrace{\mathbb{H}[\mathbf{y}_*|\mathbf{x}_*, D]}_{\text{predictive uncertainty}} = \underbrace{\mathbb{I}[\mathbf{y}_*, \mathbf{w}|\mathbf{x}_*, D]}_{\text{epistemic uncertainty}} + \underbrace{\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|D)}[\mathbb{H}[\mathbf{y}_*|\mathbf{x}_*, D]]}_{\text{aleatoric uncertainty}} \tag{2.24}$$

For instance, predictive uncertainty for a supervised classification problem is given by [DHLDVU18,

HHGL11, Cha18, GIG17, Gal16]:

$$\mathbb{H}[\mathbf{y}|\mathbf{x}, D] = -\sum_{c=1}^{C} p(\mathbf{y}|\mathbf{x}, D) log p(\mathbf{y}|\mathbf{x}, D) \tag{2.25}$$

$$= -\sum_{c=1}^{C} \Big(\int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w} \Big) log \Big(\int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w} \Big) \tag{2.26}$$

$$\approx -\sum_{c=1}^{C} \Big(\int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) q(\mathbf{w}) d\mathbf{w} \Big) log \Big(\int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) q(\mathbf{w}) d\mathbf{w} \Big) \tag{2.27}$$

$$\approx -\sum_{c=1}^{C} \Big(\frac{1}{T} \sum p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \Big) log \Big(\frac{1}{T} \sum p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \Big) \tag{2.28}$$

where $p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{exp[f_c^{\mathbf{w}}(\mathbf{x})]}{\sum_{c=1}^{C} exp(f_c^{\mathbf{w}}(\mathbf{x}))}$ is the softmax output of the network, $\mathbf{w}_t$ the t_{th} sample of $q(\mathbf{w})$ and C the number of classes.

Similarly, aleatoric uncertainty can be estimated as the expectation over the entropy of the distribution when the parameters are fixed. Analytically,

$$\mathbb{E}_{\mathbf{w} \sim p(\mathbf{w}|D)}[\mathbb{H}[\mathbf{y}|\mathbf{x}, \mathbf{w}]] = -\int p(\mathbf{w}|D) \Big[\sum_{c=1}^{C} p(\mathbf{y}|\mathbf{x}, \mathbf{w}) log p(\mathbf{y}|\mathbf{x}, \mathbf{w}) \Big] d\mathbf{w} \tag{2.29}$$

$$\approx -\int q(\mathbf{w}) \Big[\sum_{c=1}^{C} p(\mathbf{y}|\mathbf{x}, \mathbf{w}) log p(\mathbf{y}|\mathbf{x}, \mathbf{w}) \Big] d\mathbf{w} \tag{2.30}$$

$$\approx -\frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{C} p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) log p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \tag{2.31}$$

Finally, epistemic uncertainty, which originates from the model ignorance of the data distribution, can be estimated as the mutual information of $\mathbf{y}$ and $\mathbf{w}$:

$$\mathbb{I}[\mathbf{y}, \mathbf{w}|\mathbf{x}, D] = \mathbb{H}[\mathbf{y}|\mathbf{x}, D] - \mathbb{E}_{\mathbf{w} \sim p(\mathbf{w}|D)}[\mathbb{H}[\mathbf{y}|\mathbf{x}, \mathbf{w}]] \tag{2.32}$$

$$\approx -\sum_{c=1}^{C} \Big(\frac{1}{T} \sum_{t=1}^{T} p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \Big) log \Big(\frac{1}{T} \sum_{t=1}^{T} p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \Big) \tag{2.33}$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{C} p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) log p(\mathbf{y}|\mathbf{x}, \mathbf{w}_t) \tag{2.34}$$

Calibration in Deep Neural Networks Although the recent progress in Deep Learning boosted the performance in different machine learning tasks, as it is proved in [GPSW17] the DNNs output is miscalibrated in many cases. Consequently, it is very important to have

a calibrated confidence measure along with the network's prediction. Post-hoc calibration methods have been proposed without modifying the training procedure. These methods provide a mapping function from the model's output(logits) to a new probability which better estimates the actual confidence and provide a measure for uncertainty over the output of the network. Suppose a model (e.g DNN) predicts a class $\hat{Y}$ with associated confidence $\hat{P}$ and the model's logits $\mathbf{z}$. The model is calibrated if $\hat{p}$ is always the true probability, i.e $P\big(\hat{Y} = y|\hat{P} = p\big) = p$ for $p \in [0, 1]$ and class labels $y = \{1, ..., K\}$ where K is the number of classes [GPSW17]. The difference between the two sides of the above equation is defined as the expected calibration error (ECE), i.e $ece = \mathbb{E}_{\hat{P}}\big[|P(\hat{Y} = y|\hat{P} = p) - p|\big]$.

One of the most well-known calibration methods in DNNs is the one proposed in [GPSW17, LZWJ20] (temperature scaling) and it is based on parametric approach of Platt scaling [Pla99]. Let $\mathbf{z_i}$ be the logits of the model, the $\hat{p}_i$ is computed as:

$$\hat{p}_i = \max \sigma_{sm}(\mathbf{z_i^k}) \text{ where } \sigma_{sm}(\mathbf{z}_i^k) = \frac{exp(z_i^k)}{\sum_{k=1}^{K} exp(z_i^k)}$$

Using a scalar parameter $T > 0$, the confidence prediction will be:

$$\hat{q}_i = \max_k \sigma_{sm}(\mathbf{z}_i/T)^k$$

where $k = \{1, ..., K\}$ is the class label. Temperature T is optimised using negative log-likelihood on the validation set.

A more recent calibration technique is presented in [KPNK$^+$19], called Dirichlet calibration. The Dirichlet calibration method is based on Beta-calibration (for binary classifiers) [KSFF17], however this approach is applied to multi-class classifiers.

Deep Ensembles were proposed as an alternative to Bayesian NNs (e.g variational inference, MCMC) by [LPB17]. The key idea is to retrain multiple neural networks with different initializations (and random shuffling). An ensemble is treated as:

$$p(y|x) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta_m}(y|x, \theta_m)$$

i.e averaging the predicted probabilities. M is the number of networks and $\{\theta_m\}_{m=1}^{M}$ are the

parameters of the ensemble. Predictive uncertainty can be estimated using the entropy of the predictive distribution.

2.2.5 Uncertainty Estimation in Medical Imaging

Estimation of epistemic as well as aleatoric uncertainty has been widely used in medical image analysis, in different tasks such as classification, segmentation, regression, reconstruction and localisation. Some works focus solely on the estimation of aleatoric or epistemic uncertainty, while in other works authors try to estimate both aleatoric as well as epistemic uncertainty.

Monte-Carlo Dropout (MC-Dropout), as was described above, is a well-accepted approach to quantifying uncertainty [GG16b]. Several studies proposed MC-Dropout for estimation of uncertainty in different stages of the medical imaging processing pipeline. In many works, MC-Dropout is applied to a classification task [LIKO19, GGG+19, LIO19, LASA+17, CZSB20]. In [LIKO19] authors modify a ResNet [HZRS16] architecture. Dropout is added before every building block of a ResNet. Monte-Carlo dropout is utilised at test time in order to estimate epistemic uncertainty. Multiple forward passes are performed to get a distribution of the class labels. Entropy and variance of distribution are utilised for estimation of uncertainty. Leibig et al. [LASA+17] utilised MC-Dropout at test time in order to estimate uncertainty for detection of diabetic retinopathy in fundus images. Furthermore, in [LIO19] the authors proposed MC-Dropout at test time as well as a variational inference approach (i.e predicting the parameters of the posterior distribution) for predictive uncertainty (variance of the posterior). The method was applied to ambiguous classification of OCT-scans. In [CZSB20] Monte Carlo (MC) integration is applied on two popular CNNs models, DenseNet-121 and ResNet-121, for five class polyp classification. Additionally, confidence calibration using temperature scaling is applied. A comparison of uncertainty measures using Bayesian deep learning is attempted in [FFG+19]. Using variants of VGG-like models, they apply different methods such as deep ensembles, ensemble MC-Dropout, MC-Dropout and Mean-Field Variational inference to different tasks, namely out-of-data distribution detection and robustness to distribution shift, in diabetic retinopathy. In a classification and active learning framework of histopathological images of cancer tissue samples, Variational Dropout has also been applied in [RMZS19]. For uncertainty prediction, authors apply variational dropout [GG16a] and implement two types of

uncertainty, namely entropy and BALD [HHGL11, GIG17].

Uncertainty estimation based on MC-Dropout has also been applied to segmentation [RCN+19, NPAA18, OSB+19, ERVOC19, JR19, dSPBC19, HAB+20, LDW+20, NKK20, WKJ18]. In [WKJ18] MC-Dropout is applied to three Fully Convolutional Networks (FCNs) (variants of SegNet [BKC17], FCN-8 [SLD17] and U-Net [RFB15]) for semantic segmentation of colorectal polyps. Guided Backpropagation is utilised for visualisation purposes.

In [RCN+19] authors add dropout layers after every encoder and decoder block in a fully convolutional neural network (FCNN) in order to generate Monte Carlo samples for whole-brain segmentation. The samples are used to estimate different measures of structure-wise uncertainty, namely variation of the volume across the MC samples, the overlap between samples, the intersection over overlap (IOUs) metric over all MC samples. Voxel-wise uncertainty is also estimated (entropy over all MC samples). Similarly, in [NPAA18] authors train a 3D convolutional neural network with dropout and define various measures of uncertainty for multiple sclerosis lesion detection and segmentation. The uncertainty measures include the MC samples variance, predictive entropy, mutual information and prediction variance. In another work [SAR+19] uncertainty guided semi-supervised segmentation on retinal layers in oct-images is attempted. A teacher network is trained using Bayesian Deep learning (dropout in test phase) to produce soft labels as well as a confidence map, while the student model is trained taking as input the outputs of the teacher model. They also propose a novel loss function. Additionally, in [SMAN19] CNN with dropout layers is used for uncertainty quantification. Authors also applied a semi-supervised learning model, a Graph Convolutional Network (GCN), using the high confidence voxels in order to refine the output of the model.

U-Net [RFB15, ÇAL+16] becomes a very popular network for medical image segmentation. Adding dropout in its layers in order to estimate epistemic uncertainty is widely used. As an example in [OSB+19] authors compute epistemic uncertainty by adding dropout after each convolutional block of a modified U-Net network for segmentation in pathological OCT scans. Similarly in [HAB+20] a U-Net along with dropout is utilised in order to estimate uncertainty in lung nodule segmentation. Comparison with deep ensembles of U-Nets is also considered. Variations of U-Net networks, such as Dense-UNet (U-Net with DenseNet121 as encoder), Res-UNet (U-Net with ResNet as encoder) and a Res-UNet without skip or residual connections,

along with test-time dropout for uncertainty estimation is also proposed in [NKK20] for MRI brain tumor segmentation. A variant of a U-Net model, Bayesian U-Net (MC-dropout) is also proposed for uncertainty quantification in a segmentation task [HOT+19].

A U-Net together with MC-Dropout as an uncertainty measure was also proposed in [dSPBC19]. Uncertainty was used as an indicator of samples that are candidates for annotation in a deep active learning framework. This method was applied to spinal cord and brain microscopic histology images for performing myelin segmentation. In [JR19] a comparison of uncertainty measures using a U-Net for brain image (tumor) and skin image (lesion) segmentation is performed. Uncertainty measures include softmax Entropy, MC-Dropout, Ensembles, aleatoric uncertainty and an auxiliary network to predict voxel-wise uncertainty. Bayesian Unet (MC-Dropout) and Bayesian CNN (MC-Dropout) are also applied in [OWB17] for pulmonary nodule detection. Finally, a multi-task U-Net based architecture is proposed in [ERVOC19] for segmentation and regression in histopathological cell counting and white matter hyperintensity counting. MC-Dropout (as well as M heads [LPC+15]) is utilised in both segmentation of images as well as in regression.

MC-Dropout for quantification of uncertainty has also been proposed in other tasks such as medical image translation [RHC+20], where a modified version of a U-Net along with MC-dropout is applied in a CT-to-MR image translation task, in order to estimate both epistemic and aleatoric (prediction variance [KG17]) uncertainty.

In a few works, MC-Dropout is combined with test data augmentation in order to estimate uncertainty. For instance, in [CHP+20] test time MC-Dropout is applied for epistemic uncertainty estimation and a test data augmentation method is applied for estimation of aleatoric uncertainty for a skin lesion classification task. Furthermore, in [VPNN20] a novel uncertainty-based data selection scheme for omni-supervised learning, based on a 3D U-Net, is proposed. The model's uncertainty is computed using MC-Dropout at test time, while aleatoric uncertainty is computed using data augmentation (translation, scaling, rotation, reflection) at test time. The method is applied to MRI and ultrasound data segmentation. Combination of MC-Dropout and data augmentation at test phase is also applied in [LCED20] for medical image localisation. Authors propose a two-stage learning framework based on a U-Net model. Initially, the network derives a segmentation. After a post-process stage, another 3D U-Net is utilised for

the per-voxel regression of the cropped volume from the previous stage, in order to localise the Anterior Nucleus of the Thalamus. Epistemic uncertainty is computed using MC-Dropout while aleatoric uncertainty using test data augmentation. Furthermore, a novel metric called Maximum Activation Dispersion is proposed. It measures the consistency of the maximum activation positions of the Monte Carlo samples and ignores the activation variance at the same position.

Estimation of uncertainty in regression tasks, such as the prediction of disease progression is also attempted in [TLP+19]. A 3D CNN with MC-Dropout is applied to predict the progress of multiple sclerosis disease in patients, based on MRI data, within one year from the baseline scan. An existing issue is the miscalibrated predictive uncertainty in deep neural networks for regression [LIF+20]. In order to tackle this issue, authors in [LIF+20] proposed a method to calibrate the uncertainty that is derived in regression tasks. The method is based on the adjustment of the variance in the likelihood model by a trainable scalar factor.

A modified Bayesian Neural Network (BNNs) method is applied in [TKG21] for breast histopathological images classification. Authors introduce a learnable activation function that adapts to the training data. Also, they quantify the uncertainty of the predictions using as uncertainty measure the variance of the predictive probability distribution. BNNs in segmentation have been used in many works. In [KW+20] uncertainty quantification, both aleatoric and epistemic, is computed using BNNs and variational inference in an ischemic stroke lesion segmentation dataset and digital retinal images. A Bayesian residual U-Net along with a combination of dropout and DropConnect methods (DropWeights) for uncertainty quantification is proposed in [GTS20] for nuclei image segmentation. Segmentation using 3D BNNs is also applied in [LMR19] for credible geometric uncertainty on CT scans of graphite electrodes and laser-welded metals. A review of bayesian models for uncertainty estimation of imaging biomarkers and segmentation of liver in patients with diabetes mellitus is presented in [SGRP+20]. A review of uncertainty methods, including Bayes by Backprop, MC-Dropout and deep ensembles in Cardiac MRI Segmentation is also presented in [NGB+20]. For regression tasks such as bone age prediction, Bayes by Backprop [BCKW15] has been proposed [ECPUS19] in 3D MRI images.

Deep Ensembles [LPB17] is widely used for estimation of uncertainty. Although deep ensem-

bles have been mainly used for boosting the performance (e.g classification performance), they proved very useful for estimation of predictive uncertainty. The main idea is to train different networks in the whole training dataset using random initialisation and random shuffling of data samples. For the classification task, the final prediction corresponds to averaging the predicted probabilities of each network. For M models/networks the final prediction would be: $p(y|x) = \frac{1}{M}\sum_{m=1}^{M} p_{\theta_m}(y|x,\theta_m)$, where $p_{\theta_m}(y|x,\theta_m)$ is the predicted probabilities of each network separately. This method of uncertainty quantification has been applied in medical imaging, in various works such as [MWIT$^+$20, JR19, FFG$^+$19]. In [MWIT$^+$20] an ensemble of K FCNs is proposed for confidence calibration and the performance in out-of-distribution test subjects is examined. The above method is applied to brain, heart ventricle and prostate segmentation tasks. Deep Ensembles and MC-Dropout are also utilised for uncertainty estimation in [XCLT19] with application to phase imaging.

Other approaches have also been proposed for uncertainty estimation in medical imaging. In [ARA$^+$16] an uncertainty quantification method is introduced for brain segmentation using a Conditional Random Field with maximum a posteriori random perturbation. A data-driven approach for quantification of uncertainty is utilised in [AB18, AKA$^+$20] where authors use data augmentation at test time to capture heteroscedastic aleatoric uncertainty in the diabetic retinopathy detection problem. Data augmentation methods include geometric and color transformations at test time. Also, in [WLA$^+$19] authors propose a test-time, augmentation-based aleatoric uncertainty method. They examine the effect of explicitly modeled spatial transformations of the input image as well as the effect of adding noise to the input image, in the segmentation result during the test phase. MC sampling was utilised to estimate the distribution of the segmentation output. They also compare the aleatoric uncertainty with the model's uncertainty (test-time dropout). They applied their method in 2D-3D MRI brain image segmentation.

In [NEN$^+$19] authors propose a framework to quantify uncertainty in left ventricle segmentation task. Spatial transformations are applied on each input image. The statistical variance of the network's output is defined as an indicator of the network uncertainty. After computing the above deviation for each pixel, authors apply an adaptive thresholding algorithm based on Conditional Random Fields (CRFs) to obtain the final segmentation. In [DLX$^+$20] a model-

independent approach for quantification of uncertainty in selective segmentation is proposed. In this direction, they propose a novel uncertainty function in the training phase. The method was evaluated using different datasets for whole heart segmentation and for gland segmentation.

Evidence theory based uncertainty quantification is applied in [XYLD20]. Authors utilise Dempster-Shafer evidence theory to estimate the uncertainty in deep neural networks in a binary classification problem on chest X-Rays and breast images. Furthermore, in [GGG+19] authors proposed an explicit mechanism to learn classification uncertainty in order to reject samples with high uncertainty. The method was based on the Dempster-Shafer framework for modeling of evidence [Dem68]. The method was applied to thoracic diseases classification.

Generative Adversarial Networks (GANs) have also been proposed for deep active learning in a semi-supervised setting using uncertainty information as an indicator for the choice of sample which will then will be sent to an annotator. For instance, in [RKBN19] authors suggest the use of a conditional GAN (cGAN) model and utilise the discriminator output score as an indicator for the uncertainty of each sample. The method is evaluated on 3D cardiovascular MR images.

It is proved that quantification of uncertainty could be also useful in domain adaptation problems, especially in cases for which the amount of labeled training data is limited. For instance in [CGP+20] inherited uncertainty is exploited in order to improve the quality of segmentations in the target domain images. The method was evaluated in MRI prostate cancer images. For domain alignment, an uncertainty-aware cross entropy loss for anatomical structure segmentation is also proposed in [BYW+20].

In [ZPP+20] a FCN is applied for cartilage segmentation in 3D mirco-CT images. In order to estimate uncertainty, a bootstrap ensemble based uncertainty quantification method is proposed. The framework consists of a FCN which is trained to predict pseudo-labels and uncertainty maps for unseen slices. Then, guided by the uncertainty, the FCN is trained using the pseudo-labels in order to improve the generalization ability of the FCN. Furthermore, the FCN(s) is integrated into a bootstrap ensemble and a K-head FCN is devised.

Furthermore, an uncertainty-guided loss method using a self-supervised task for updating a FCN model is applied in [LCX+20].

A double-uncertainty weighted method based on the teacher-student model is introduced in [WZT+20],

where double-uncertainty is used to guide the teacher model. Furthermore, an online training uncertainty loss is also proposed in [XDS+19] where authors suggest defining a sample as certain/uncertain based on the training loss and subsequently only samples with low loss contribute to the backpropagation process. Furthermore, they use a re-weighting method which is based on the probabilistic Local Outlier Factor (pLOF) in order to preserve the influence of the minority class. The method is evaluated in skin lesion classification.

Uncertainty quantification has also been examined in medical image synthesis tasks [UVA20]. Authors trained a GAN-based network using quasi-norm based penalties to synthesize T2w from T1w brain MRI images. Estimation of uncertainty in other tasks of the medical imaging pipeline is attempted, such as quality transfer/classification [TWK+21, TWG+17] and MRI reconstruction in k-space [ZRM+19].

For uncertainty quantification in segmentation, probabilistic models are proposed in [KRPM+18, BKCTC+19, GEK20]. These works are focused on the generation of diverse/plausible segmentations from a single or multiple annotations and the connection of these outputs with the inter-observer variability.

In [KRPM+18] the proposed method (combination of a U-Net [RFB15] with a conditional variational autoencoder) minimises the Kullback-Leibler divergence between a prior and a posterior network. A more detailed description of this method will be given in Chapter 4. Extensions of this method were presented in [BKCTC+19] and in [GEK20]. In [BKCTC+19] a hierarchical probabilistic model is proposed. In this model, latent variables are utilised in order to model the segmentation at different resolutions. Finally, in [GEK20] authors proposed reversible blocks [GRUG17], modifying the previous architecture in order to alleviate the issue of suffering from a significant memory burden during training.

Furthermore, a use case of uncertainty estimation is to provide insight for unusual patterns, i.e anomalies (or novelty) and out-of-distribution data. In medical imaging most works of this type focus on the segmentation task. MC-Dropout is applied in a U-Net for CT image translation to MRI for unsupervised anomaly segmentation in [RHH+20]. They utilise as anomaly score the division of epistemic with aleatoric uncertainty and they define as anomalies voxels, the voxels with high value in this fraction. In [SOS+20] a Bayesian U-Net is applied for segmentation using weak labels which are automatically generated with a graph-based segmentation approach. In

the next step MC-Dropout is applied in order to compute epistemic uncertainty. The output of the network is thresholded in order to highlight uncertain/anomalous regions. Finally, a post processing method is applied in order to derive the final segmentation map. Unsupervised anomaly segmentation in head MRI is also attempted in [SHMU19], where uncertainty-based anomaly score is derived through a modification of variational autoencoder's loss function.

A summary of uncertainty quantification studies applied to medical imaging is provided in Table 2.5.

Reference	ML task	Method	Application Domain
[MWIT+20]	Segmentation	Deep Ensembles	Brain, heart ventricle, Prostate (MR)
[FFG+19]	Classification	Deep Ensembles MC-Dropout BNNs-MFVI	fundus images Diabetic Retinopathy
[LASA+17]	Classification	MC-Dropout	fundus images Diabetic Retinopathy
[GGG+19]	Classification	method based on Dempster-Shafer theory MC-Dropout	Chest radiograph assessment
[LIO19]	Classification	MC-Dropout Variational Inference (ResNet)	OCT-scans
[CZSB20]	Classification	MC integration	Polyp-Colonoscopy Images
[RCN+19]	Segmentation	MC-Dropout	MRI scan Brain
[NPAA18]	Detection Segmentation (lesion)	MC-Dropout	multiple sclerosis MR Brain
[OSB+19]	Segmentation	MC-Dropout	OCT scans (pathological)
[HAB+20]	Segmentation (nodules)	Deep Ensembles MC-Dropout	lung CT scans
[NKK20]	Segmentation (tumors)	MC-Dropout	Brain MRI
[dSPBC19]	Deep Active Learning (Axon-Myelin Segmentation)	MC-Dropout	Histology Data (spinal cord & Brain)
[JR19]	Segmentation (tumors & lesions)	Deep Ensembles MC-Dropout Aleatoric Uncertainty Auxiliary network (to predict uncertainty)	MRI Brain skin imaging
[ERVOC19]	Segmentation Regression (cell counting, lesion counting)	MC-Dropout	Cell histology & White Matter lesions
[RHC+20]	Image translation (from CT to non-contrast T1-w MRI)	MC-Dropout aleatoric (prediction variance)	Brain

[TLP+19]	Regression (prediction of disease progression)	MC-Dropout	MRI Brain
[VPNN20]	Segmentation	MC-Dropout test data augmentation	MRI brain & Ultrasound (neurosonography)
[LCED20]	Localisation (Anterior Nucleus of the Thalamus)	test data augmentation MC-Dropout Maximum Activation Dispersion	MRI brain
[CHP+20]	Classification (skin lesion)	MC-Dropout test data augmentation	skin dermoscopic images
[TKG21]	Classification	modified BNNs	breast histopathological images
[RMZS19]	Classification Active Learning	Variational Dropout	histopathological Hematoxylin-Eosin (H & E) (colorectal cancer)
[KW+20]	Segmentation	BNNs-VI	MRI brain & colored retina images
[SAR+19]	Segmentation	BNNs (dropout) (teacher-student models)	OCT data
[HOT+19]	Segmentation	Bayesian U-Net (MC-dropout)	CT scan hip (arthroplasty & soft tissue sarcoma)
[OWB17]	Detection	Bayesian U-Net & CNN (MC-Dropout)	CT scan (pulmonary nodules)
[GTS20]	Segmentation Detection	Bayesian Res-UNet/VI (DropWeights)	microscopy images (nuclei images) MRI brain
[ECPUS19]	Regression (bone age prediction)	Bayes by Backprop	T1-w MR images (wrists and clavicles)
[LMR19]	3D segmentation	3D CNN-VI	CT scans (graphite electrodes and laser-welded metals)
[NEN+19]	Segmentation	(statistical) variance of random transformations of input images	Cardiac MRI (left ventricle)
[AB18] [AKA+20]	Classification	test time data augmentation	fundus images diabetic retinopathy
[DLX+20]	Segmentation	plug-and-play method introducing an uncertainty loss	heart & gland
[WKJ18]	Segmentation	MC-Dropout	colorectal polyps (colonoscopies)

[XYLD20]	Classification (binary)	Evidential DNNs	Chest X-Rays Breast
[LIF+20]	Regression (calibration)	MC-Dropout Aleatoric	Bone (CT) OCT (6DoF needle pose) breast (H & E stained) invasive surgery (endoscopic images)
[RKBN19]	Segmentation Deep Active Learning	GAN-based (discriminator output)	MR cardiac images
[SMAN19]	Segmentation	MC-Dropout (+ GCN refinement)	CT images pancreas & spline
[LCX+20]	Segmentation	self-loop uncertainty	H&E stained tissue (different organs) skin lesion images
[CGP+20]	Segmentation Domain adaptation	GAN-based method	MRI prostate (cancer)
[WLA+19]	Segmentation	test data augmentation MC-dropout	MRI Brain
[BYW+20]	Segmentation Domain adaptation	Adversarial learning uncertainty-aware loss	OCT (retinal) & MRI/CT (cardiac)
[ZPP+20]	Segmentation (cartilage)	bootstrap ensemble	micro-CT images
[WZT+20]	Segmentation	feature uncertainty uncertainty consistency loss MC-Dropout	gadolinium-enhanced MRI (left atrium) & abdominal CT (kidney)
[XDS+19]	Classification	online uncertainty sample mining method	skin lesion
[LDW+20]	Segmentation	Correlation-based	cardiac magnetic resonance (ventricle)
[ARA+16]	Segmentation (tumours)	Conditional Random Field (random perturbation)	MR Brain
[UVA20]	image synthesis	GAN-based (quasi-norm penalties)	MR Brain (T2w from T1w)
[TWK+21] [TWG+17]	Classification Quality transfer	heteroscedastic noise variational dropout	diffusion MRI (brain) (Neuroimamge)
[XCLT19]	phase imaging	(U-Net) Deep Ensembles MC-Dropout	biological samples
[ZRM+19]	MRI Reconstruction (k-space)	Bayesian deep learning	knee (DICOM)

[BKCTC+19]	Segmentation	Variational Inference-based hierarchical probabilistic model	CT (lung) MR (prostate)
[GEK20]	Segmentation	Variational Inference-based hierarchical probabilistic model (reversible blocks)	CT (lung) MR (prostate)
[KRPM+18]	Segmentation	Variational Inference-based U-Net with cVAE	CT scan (lungs)
[RHH+20]	Segmentation (anomaly detection)	MC-Dropout aleatoric uncertainty (predicted variance)	MRI & CT brain
[SOS+20]	Segmentation (anomaly detection)	MC-Dropout (Bayesian U-Net)	retinal OCT scans
[SHMU19]	Segmentation (anomaly detection)	Variational Inference Variational Autoencoder	head MRI head
[SGRP+20]	Segmentation	MC-Dropout Bayesian FCNN Probabilistic U-Net Hierarchical Probabilistic U-Net	MRI
[LIKO19]	Classification	MC-Dropout	Retinal OCT scan

Table 2.5: A review of studies of Uncertainty Quantification in Medical Imaging

2.3 Conclusion

This Chapter covered extensively the concepts of Anomaly Detection as well as Uncertainty Estimation in Medical Imaging. Both research areas present rapidly growing literature, especially in the context of the recent advances in deep learning. However, there are still challenges and potential directions for future research. A comparative experimental analysis of the existing methods, both in medical imaging anomaly detection as well as in Uncertainty Estimation in medical imaging, should be attempted. A promising research avenue which should be further explored is the anomaly detection based on uncertainty. Furthermore, model interpretability has not received so much attention in existing works, in both research areas.

Chapter 3

Anomaly Detection in Fetal Screening

In this Chapter, an automated framework for detection of cardiac anomalies during ultrasound screening is proposed and evaluated on the example of Hypoplastic Left Heart Syndrome (HLHS), a sub-category of congenital heart disease. Congenital heart disease is the most common group of congenital malformations, affecting $6 - 11$ per 1000 newborns. An unsupervised approach, that learns healthy anatomy exclusively from clinically confirmed normal control patients, is proposed. We evaluate a number of known anomaly detection frameworks together with a model architecture based on the α-GAN network and find evidence that the proposed model performs better than the state-of-the-art in image-based anomaly detection, yielding average 0.81 AUC and a better robustness towards initialisation compared to previous works.

3.1 Introduction

A contemporary key element in building automated pathology detection systems with machine learning in medical imaging is the availability and accessibility of a sufficient amount of data in order to train supervised discriminator models for accurate results. This is a problem in medical imaging applications, where data accessibility is scarce because of regulatory constraints and economic considerations. To build truly useful diagnostic systems, supervised machine learning methods would require a large amount of data and manual labelling effort for every possible disease to minimise false predictions. This is unrealistic because there are thousands of diseases, some represented only by a few patients ever recorded. Thus, learning

representations from healthy anatomy and using anomaly detection to flag unusual image features for further investigation defines a more reasonable paradigm for medicine, especially in high-throughput settings like population screening, e.g. fetal ultrasound imaging. However, anomaly detection suffers from the great variability of healthy anatomical structures from one individual to another within patient populations as well as from the many, often subtle, variants and variations of pathologies. Many medical imaging datasets, e.g. volunteer studies like UK Biobank [PMB+13], consist of images from predominantly healthy subjects with a small proportion of them belonging to abnormal cases. Thus, an anomaly detection approach or 'normative' learning paradigm is also reasonable from a practical point of view for applications like quality control within massive data lakes.

In this Chapter, we formulate the detection of congenital heart disease as an anomaly detection task for fetal screening with ultrasound imaging. We utilise normal control data to learn the normative feature distribution which characterises healthy hearts and distinguishes them from fetuses with hypoplastic left heart syndrome (HLHS). We chose this test pathology because of our access to a well labelled image database from this domain. Theoretically, our method could be evaluated on any congenital heart disease that is visible in the four-chamber view of the heart [NHS15].

3.1.1 Pathological Diseases in Fetal Heart

Congenital heart disease (CHD) is the most common group of congenital malformations [BMG+10, YLR18, vVCR+16]. CHD is a defect in the structure of the heart or great vessels that is present at birth. Approximately $6-11$ per 1000 newborns are affected [vVCR+16]. $20-30\%$ of these heart defects require surgery within the first year of life [vVCR+16]. In order to detect the disease, the most common approach is the standard anomaly ultrasound scan at approximately 20 weeks of gestation (e.g. 18+0 to 20+6 weeks in the UK). In contemporary screening pathways, *i.e.*, 2D ultrasound at GA 12 and 24, the prenatal detection rate of CHD is in a range of $39-59\%$ [PKM+12, vVCR+16]. In [YLR18], algorithmic support has been used to find diagnostically informative fetal cardiac views. With this aid, clinical experts have been shown to discriminate healthy controls from CHD cases with 98% sensitivity and 93% specificity in 4D ultrasound. However, 4D ultrasound is not commonly used during fetal screening and in the

proposed teleradiology setup still all images have to be manually assessed by highly experienced experts to achieve such a high performance.

In this Chapter, we focus on a subtype of CHD, Hypoplastic Left Heart Syndrome (HLHS). Examples of HLHS in comparison with healthy fetal hearts are presented in Figure 3.1. HLHS is rare, but is one of the most prominent pathologies in our cohort. In HLHS the four chamber view is usually grossly abnormal, allowing the identification of CHD (although not necessarily a detailed diagnosis) from a single image plane. A condition that is identifiable on a single view plane provides a clear case study for our proposed method. If HLHS is identified during pregnancy, provisions for the appropriate timing and location of delivery can be made, allowing immediate treatment of the affected infant to be investigated after birth. Postnatal palliative surgery is possible for HLHS, and the antenatal diagnosis of CHD in general has been shown to result in a reduced mortality compared to those infants diagnosed with CHD only after birth [HMWJ15]. However, the detection of this pathology during routine screening still remains challenging. Screening scans are performed by front-line-of-care sonographers with varying degrees of experience and the examination is influenced by factors such as fetal motion and the small size of the fetal heart.

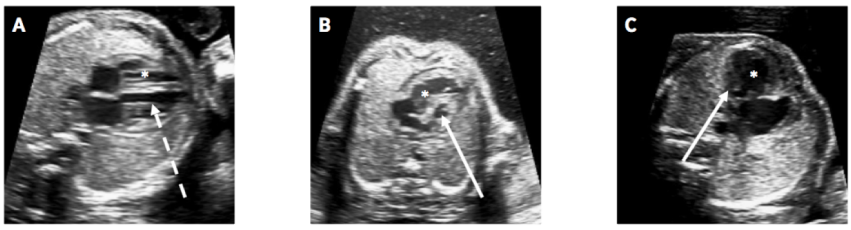


Figure 3.1: Examples of four-chamber views of the fetal heart. A shows a normal fetal heart, with the normal sized LV (left ventricle) marked (dashed white arrow). B and C show two examples of fetal HLHS (hypoplastic left heart syndrome), with the hypoplastic LV marked (solid white arrow). Example B represents the mitral stenosis / aortic atresia subtype, with a severely hypoplastic, globular LV. Example C represents the mitral atresia / aortic atresia subtype, with a slit-like LV that is difficult to identify. * marks the right ventricle in each case.

3.1.2 One-class anomaly detection methods in Medical Imaging

In one-class classification, (training) data from only a single class are available [PP18]. The main goal is to learn either a representation or a classifier (or a combination of both) in order

to distinguish and recognise out-of-distribution samples during inference. Discriminative as well as generative methods have been proposed utilizing deep learning, for example one class CNN [OP19] and Deep SVDD [RVG+18]. Usually these methods utilise loss functions, similar to those of OC-SVM [SPST+01] and SVDD [TD04] or use regularisation techniques to make conventional neural networks compatible to one-class classification models [POP21].

Our approach attempts to model the distribution of healthy data. Towards this direction, deep generative models, such as generative adversarial networks and variational autoencoders have achieved state-of-the-art performance in modeling complicating distributions [ZCLZ16, SSW+19, SBW21]. Furthermore, generative models (i.e Generative Adversarial Networks) show great success in generating natural-looking images [RSDM19, SBW21]. This is very important for our approach, which benefits from the use of the reconstruction component as well. Consequently, in this Chapter, we focus on the application of generative models via an adversarial process for (one-class) anomaly detection.[1]

Generative adversarial networks for anomaly detection were first proposed by [SSW+17]. In [SSW+17], a deep convolutional generative adversarial network, inspired by DCGAN as proposed by [RMC16], is used as *AnoGAN*. During the training phase, only healthy samples are used. This approach consists of two models. A generator, which generates an image from random noise and a discriminator, which classifies real or fake samples as common in GANs. More specifically, the generator learns the mapping from the uniformly distributed input noise sampled from the latent space to the 2D image space of healthy data. The output of the discriminator is a single value, which is interpreted as the probability of an image to be real or generated by the generator network. In their work, a residual loss is introduced, which is defined as the $l1$ norm between the real images and the generated image. This enforces the visual similarity between the initial image and the generated one. Furthermore, in order to cope with GAN instability, instead of optimizing the parameters of the generator via maximizing the discriminator's output on generated examples, the generator is forced to generate data whose intermediate feature representation of the discriminator (D_H) is similar to those of real images [SSW+17]. This is defined as the $l1$ norm between intermediate feature representations of the discriminator given as input the real image and the generate image respectively.

[1] An overview of deep learning-based, medical AD methods was given in Section 2.1.3.

In AnoGAN, an anomaly score is defined as the loss function at the last iteration, *i.e.*, the residual error plus the discrimination error. AnoGAN has been tested on a high-resolution SD-OCT dataset. For evaluation purposes, the authors report receiver operating characteristic (ROC) curves of the anomaly detection performance. Based on their results, using the residual loss alone already yields good results for anomaly detection. The combination with the discriminator loss (i.e adversarial score) improves the overall performance slightly. During testing, an iterative search in the latent space is used in order to find the closest latent vector that generates an image that is the most similar to the real test image. This is a time consuming procedure and this optimisation process can get stuck in local minima.

Similar to AnoGAN, a faster approach, f-AnoGAN has been proposed in [SSW+19]. In this work, the authors train a GAN on normal images, however instead of the DCGAN model a Wasserstein GAN (WGAN) [ACB17] [GAA+17] has been used. Initially, a WGAN is trained in order to learn a non-linear mapping from latent space to the image space domain. Generator and discriminator are optimised simultaneously. Samples that follow the data distribution are generated through the generator, given input noise sampled from the latent space. Then an encoder (convolutional autoencoder) is training to learn a map from image space to latent space. For the training of the encoder, different approaches are followed, i.e training an encoder with generated images (z-to-z approach-*ziz*), training an encoder with real images (an *image to latent space to image* mapping approach -*izi*) and training a discriminator guided izi encoder (izi_f). As anomaly score, image reconstruction residual plus the residual of the discriminator's feature representation (D_H) is used. The method is evaluated on optical coherence tomography imaging data of the retina. Both [SSW+17] as well as [SSW+19] use image patches for training and are modular methods which are not trained in an end-to-end fashion [ZGC+20].

Another GAN-based method applied to OCT data has been proposed by [ZGC+20], in which authors propose a Sparsity-constrained Generative Adversarial Network (Sparse-GAN), a network based on an Image-to-Image GAN [IZZE17]. Sparse-GAN consists of a generator, following the same approach as in [IZZE17], and a discriminator. Features in the latent space are constrained using a Sparsity Regularizer Net. The model is optimized with a reconstruction loss combined with an adversarial loss and sparsity regularization. The anomaly score is computed in the latent space and not in image space. Furthermore, an Anomaly Activation Map (AAM) is

proposed to visualise lesions.

Subsequently, AnoVAEGAN [BWAN18] has been proposed, in which the authors discuss a spatial variational autoencoder and a discriminator. It is applied to high resolution MRI images for unsupervised lesion segmentation. AnoVAEGAN uses a variational autoencoder and tries to model the normal data distribution that will lead the model to fully reconstruct the healthy data while it is expected to fail reconstructing abnormal samples. The discriminator classifies the inputs as real or reconstructed data. As anomaly score the $l1$ norm of the original image and the reconstructed image is used.

Opposite to reconstruction-based anomaly detection methods as they are discussed above, in [SCWZ20] adGAN, an alternative framework based on GANs, is proposed. The authors introduce two key components: fake pool generation and concentration loss. adGAN follows the structure of WGAN and consists of a generator and discriminator. The WGAN is first trained with gradient penalty using healthy images only and after a number of iterations a pool of fake images is collected from the current generator. Then a discriminator is retrained using the initial set of healthy data as well as the generated images in the fake pool with a concentration loss function. Concentration loss is a combination of the traditional WGAN loss function with a concentration term which aims to decrease the within-class distance of normal data. The output of the discriminator is considered as anomaly score. The method is applied to skin lesion detection and brain lesion detection. Two other methods that utilise discriminator outputs as anomaly score, however not tested for medical imaging, are ALOCC [SKFA18] and fenceGAN [NWC+19]. In ALOCC [SKFA18], the discriminator's probabilistic output is utilised as abnormality score. In their work an encoder-decoder is used for reconstruction while the discriminator tries to differentiate the reconstructed images from the original ones. An extension of the ALOCC algorithm, is the Old is Gold (OGN) algorithm which is presented in [ZLAL20]. After training a framework similar to ALOCC, the authors finetune the network using two different types of fake images which are good/bad quality (reconstructions) images and pseudo anomaly images. In this way they try to boost the ability of the discriminator to differentiate normal images from abnormal ones.

In [NWC+19] the authors propose a modified GAN loss, so that the generated images are placed at the boundary of the normal data distribution and then use the discriminator in order

to distinguish anomalous images. They propose this loss with the idea that a conventional GAN objective encourages the distribution of generated images to overlap with real images.
In [GZZ+20] the authors proposed an adversarial one-class classification method combined with video transfer learning for the detection of fetal congenital heart disease. The proposed method consists of two parts, namely DANomaly and GACNN (Wgan-GP and CNN). The former is one-class classification approach similar to the ALOCC algorithm (using an additional noise image model) and is utilised for screening end-systolic (four chamber heart) video slices. During training both normal and abnormal samples are available (which is one of the key differences compared to our approach where only healthy subjects are utilised). In [PNX19] a one-class generative adversarial network (OCGAN) is proposed for anomaly detection. OCGAN consists of two discriminators, a visual and a latent discriminator, a reconstruction network (denoising autoencoder) and a classifier. The latent discriminator learns to discriminate encoded real images and random samples drawn from $\mathcal{U} \sim (-1, 1)$ distribution, while the visual discriminator distinguishes real from fake images. Their classifier is trained using binary cross entropy loss and learns to recognise positive examples from negative examples. Finally, in [PAD18] a probabilistic framework is proposed which is based on a model similar to α-GAN. The distribution of the latent space is forced to be similar to standard normal distribution through an extra (latent) discriminator network, similar to [RLWFM17]. A parameterized data manifold is defined (using an adversarial autoencoder) which captures the underlying structure of the inlier distribution (normal data). A test sample is considered as abnormal if its probability with respect to the inlier distribution is below a threshold. The probability is factorised with respect to local coordinates of the manifold tangent space.
A summary of the key features for the works above is given in Table 3.1.
To establish consistency between different related works we define x as a test image, $\hat{x}$ as a reconstructed image, D as a discriminator network, (D_H as (intermediate) feature representation of a Discriminator network), E as an encoder network (image space $\rightarrow$ latent space), De as a decoder network (latent space back to image space), G as a generator network (where input is a noise vector), z as latent space representation and λ as a weighting factor.

Table 3.1: One-class anomaly detection using Generative Adversarial Networks

Reference	Approach	Anomaly score	Dataset
AnoGAN [SSW+17]	residual & discrimination score	$(1-\lambda)\|x-G(z)\|+\lambda\|D_H(x)-D_H(G(z))\|$	OCT
f-AnoGAN [SSW+19]	reconstruction & discrimination score	$\|x-De(E(x))\|^2+\lambda\|D_H(x)-D_H(De(E(x)))\|^2$	OCT
Sparse-GAN [ZGC+20]	reconstruction error	$\|E(x)-E(De(E(x)))\|_2$	OCT
AnoVAEGAN [BWAN18]	reconstruction error	$\|x-De(E(x))\|_1$	Brain
adGAN [SCWZ20]	discriminator score	$D(x)$	Digit/skin/Brain
*ALOCC [SKFA18]	discriminator score	$D(De(E(x)))$	Generic Images/Video
*fenceGAN [NWC+19]	discriminator score	$D(x)$	Generic Images
*OGN [ZLAL20]	discriminator score	$D(De(E(x)))$	Generic Images/Video
*OCGAN [PNX19]	discriminator/reconstruction score	$D(De(E(x)))/\|x-De(E(x))\|^2$	Generic Images
*GPND [PAD18]	probabilistic score	$p_x(x)$	Generic Images

* Application field of these works as they are described in the original papers is not Medical Imaging.

3.2 Materials and Methods

In order to detect anomalies in fetal ultrasound data, we build an end-to-end model which takes as input the whole image and produces an anomaly score together with an attention map in an unsupervised way.

To achieve this, we build a GAN-based model, where the aim of the discriminator networks is to learn the salient features of the fetal images (*i.e.*, heart area) during training. We use an auto-encoding generative adversarial network based on (α-GAN) which makes use of discriminator information in order to predict the anomaly score. α-GAN [RLWFM17] [KHD19] is a fusion of generative adversarial learning (GAN) and a variational autoencoder. It aims to overcome GAN instabilities during training, which leads to mode collapse while at the same time exploits the advantages of variational autoencoders, producing less blurry images. In α-GAN two discriminators focus on the data and latent space respectively. An overview of the proposed architecture is given in Figure 3.2

We assume the true data distribution of real fetal cardiac images x as $x \sim p_x^*$ and a random prior distribution p_z. Reconstruction, $\hat{x}$, of an input image x is defined as $G(\hat{z})$ where $\hat{z}$ is a sample from the variational distribution q_E, *i.e.*, $\hat{z} \sim q_{E(z|x)}$. Furthermore, we define z as a sample from a normal prior distribution p_z, *i.e.*, $z \sim \mathcal{N}(0,1)$.

The encoder (E) is mapping each real data point x from image space $\mathcal{X}$ to a point in the (d-dimensional) latent space $\mathcal{Z}$, *i.e.*, $E : \mathcal{X} \to \mathcal{Z}$. It consists of four blocks. Each block contains a Convolutional-Batch Normalisation layer followed by Leaky Rectified Linear Unit (LeakyReLU) activation, down-sampling the resolution of data by two in each block. Spectral Normalisation [MKKY18, ZGMO19] a weight normalisation method, is used after each convo-

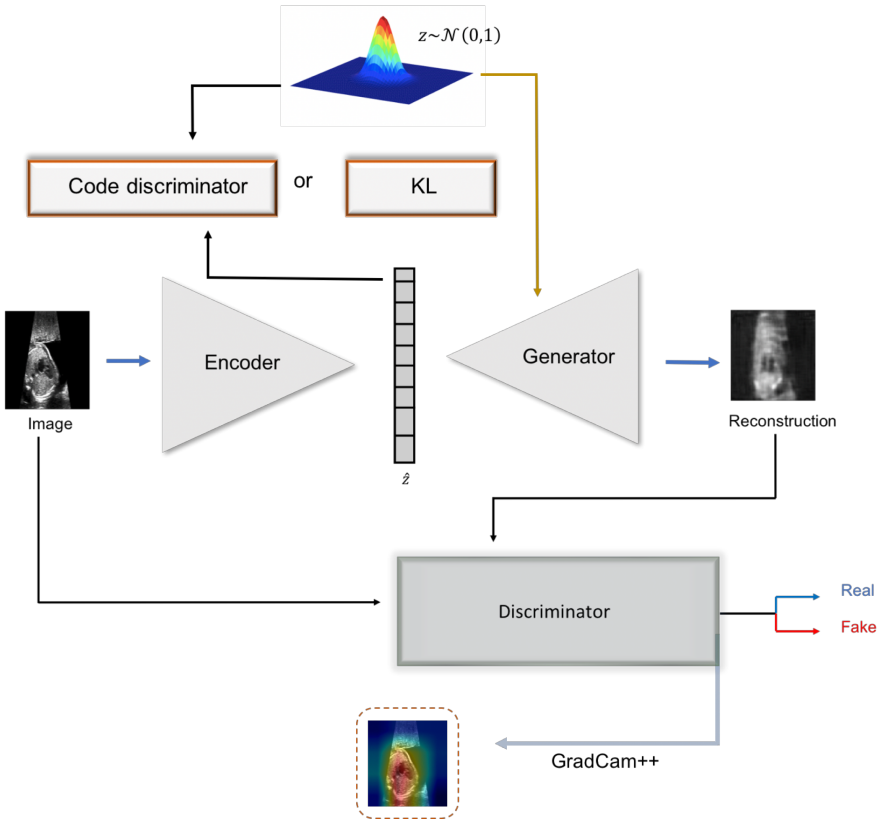


Figure 3.2: Proposed GAN-based model. The encoder is used to map the input into lower dimensional (latent) space. Generator/decoder is used either for the reconstruction of the original image from the latent space or for generating samples from a random noise vector. Additionally, two discriminators applied to image and latent space respectively are used, to distinguish real from fake samples.

lutional layer.[2] In the last block, after the convolutional block, an attention gate is introduced [SOS+19, ZGMO19]. The final layer of the encoder is a tangent layer. The dimension of the latent space is equal to 128.

The generator synthesises images from latent space $\mathcal{Z}$ back to the image space $\mathcal{X}$, *i.e.*, $G : \mathcal{Z} \rightarrow \mathcal{X}$. The generator regenerates the initial image using four consecutive blocks of transposed convolution-batch normalisation-Rectified Linear Unit (ReLU) activation layers [ZGMO19]. The last layer is a Hyperbolic tangent (tanh) activation. Similar to encoder spectral normalisation, attention gate layers are used.

The discriminator (D) takes as input an image and tries to discriminate between real and fake

[2]Spectral normalisation was proposed by [MKKY18] and normalizes the spectral norm of weight matrix W with $\frac{W}{\sigma(W)}$, where $\sigma(W)$ is the largest singular value of W. In order to compute $\sigma(W)$, power iteration method is utilised. Spectral Normalisation controls the Lipschitz constant of the discriminator function. It is used in our case as a regularisation method in order to stabilize the training procedure [MKKY18].

Algorithm 1: Training procedure of the proposed method.

Input : images x, parameter: λ, Number of Epochs: N
Output: Trained Networks

for *epoch 1 to* N **do**
  Update E, G using Eqs. 3.1, 3.2, 3.3. D, LD are fixed.
  ▷ $L_{\{.\}}$ indicates the loss function of each network
  $L_E \leftarrow \lambda \|x - \hat{x}\|_1 + LD(\hat{z}, 1)$
  $L_G \leftarrow \lambda \|x - \hat{x}\|_1 + D(\hat{x}, 1) + D(\tilde{x}, 1)$
  $L_{E,G} \leftarrow L_E + L_G$
  Update D using Eq. 3.4. E, G, LD are fixed.
  $L_D \leftarrow D(x, 1) + D(\hat{x}, 0) + D(\tilde{x}, 0)$
  Update LD using Eq. 3.5. E, G, D are fixed.
  $L_{LD} \leftarrow LD(\hat{z}, 0) + LD(z, 1)$
return G, D, LD, E
  ▷ G, D, LD, E indicates the corresponding outputs of each network and 1/0 the real/fake values

images. The output of the discriminator is a probability for the input being a real or fake image. It consists of four blocks. Each block consists of Convolutional-Batch Normalisation-RELU layers. The last layer is a sigmoid layer. The discriminator treats x as real images while the reconstruction from the encoder and samples from p_z, are considered as fake.

Instead of making a restrictive assumption, that the variational distribution $q_E(z|x)$ belongs to a particular distribution family, we can replace the KL divergence (from the evidence lower bound) with a latent code discriminator (LD) which acts in an adversarial way with E [RLWFM17].

The discriminator distinguishes the samples generated by the encoder $q_E(z|x)$ from the samples generated by the p_z(a standard Gaussian distribution). The latent code discriminator consists of four linear layers followed by a Leaky RELU activation.

We randomly initialise the encoder, generator and latent code discriminator. The weights for the discriminator are initialised with a normal distribution $\mathcal{N} \sim (0, 0.02)$. We train the architecture by first updating the encoder parameters by minimizing:

$$\mathcal{L}_{\mathrm{E}} = \mathbb{E}_{p_x^*}[\lambda \times \|x - \hat{x}\|_1 + (-log(LD(\hat{z})))] \tag{3.1}$$

We define the generator loss as:

$$\mathcal{L}_{\mathrm{G}} = \mathbb{E}_{p_x^*}[\lambda \times \|x - \hat{x}\|_1 + (-log(D(G(\hat{z}))))] + \mathbb{E}_{p_z}[-log(D(G(z)))] \tag{3.2}$$

Since we consider encoder and generator as one network the loss for the encoder-generator is:

$$\mathcal{L}_{E,G} = \mathcal{L}_E + \mathcal{L}_G \tag{3.3}$$

where $\mathcal{L}_E$ and $\mathcal{L}_G$ are defined in Eqs. 3.1 and 3.2 respectively. The generator is updating twice compared to the encoder in order to stabilize the training procedure.

Then we update (minimizing) the discriminator loss by considering as real the input images and as fake the reconstructions and the images which are derived from the generator (taking as input samples from the p_z distribution). In this way,the discriminator distinguishes between the real images and fake images generated by the network G.

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{p_x^*}\left[-2 * \log \mathrm{D}(x) - \log\left(1 - \mathrm{D}(G(\hat{z}))\right)\right] + \mathbb{E}_{p_z}\left[-\log\left(1 - \mathrm{D}(G(z))\right)\right]. \tag{3.4}$$

Finally, we update the weights of latent code discriminator using

$$\mathcal{L}_{\mathrm{LD}} = \mathbb{E}_{p_x^*}\left[-\log\left(1 - \mathrm{LD}(\hat{z})\right)\right] + \mathbb{E}_{p_z}\left[-\log(\mathrm{LD}(z))\right]. \tag{3.5}$$

For the learning rate λ, we use value of 25 after grid search.

The training process of the α-GAN model is described in algorithm 1. The networks are trained using the Adam optimizer. Encoder and Generator use the same learning rate, λ. The same learning rate is also utilised for discriminator and latent code discriminator.

We additionally replace the latent discriminator with an approximation of KL divergence. For a latent vector z of M dimension we define KL divergence as [UVL18, RLWFM17]:

$$KL(q_E(z|x)||\mathcal{N}(0, I)) \approx -\frac{M}{2} + \frac{1}{M}\sum_{i=1}^{M}\frac{s_i^2 + m_i^2}{2} - log(s_i),$$

where m_i and s_i is the mean and standard deviation of each component of the M_{th} dimensional latent space. Performance in this configuration is subpar, thus we limit the discussion to results with the latent code discriminator.

Furthermore, we apply an analytic estimation of KL divergence using a one-class variational

autoencoder (VAE-GAN) similar to [BWAN18] [DB16]. The VAE-GAN is trained using reconstruction error plus the KL divergence between the latent space ($\hat{z}$) and the normal distribution p_z. For training the VAE-GAN, we first update the encoder and decoder networks as following:

$$\mathcal{L}_E = \mathbb{E}_{p_x^*}[\beta * \|x - \hat{x}\|_p] + KL(q_E(z|x)||p_z)$$

$$\mathcal{L}_{\mathrm{G}} = \mathbb{E}_{p_x^*}[\gamma \times \|x - \hat{x}\|_p + (-log(D(G(\hat{z}))))] + \mathbb{E}_{p_z}[-log(D(G(z)))]$$

Finally, the discriminator is trained based on the:

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{p_x^*}\left[-2 * \log \mathrm{D}(x) - \log\left(1 - \mathrm{D}(G(\hat{z}))\right)\right] + \mathbb{E}_{p_z}\left[-\log\left(1 - \mathrm{D}(G(z))\right)\right].$$

where β, γ are set to 10 and 5 respectively after grid search.

A ResNet18 [HZRS16]-based architecture encoder and decoder/generator are utilised (with random initialisation). In the ResNet18 encoder/decoder architecture each layer consists of 4 residual blocks and each block is 2−layer deep. We use the same discriminator as in α-GAN. The dimensions of the latent space are 128. $p = 2$ since we use the l_2 norm (*i.e.*, mean square error). All networks are implemented in Python using Pytorch, on a workstation with a NVIDIA Titan X GPU.

3.2.1 Anomaly detection score

In order to predict an anomaly score s, three different strategies are utilised. For an unseen image x_{unseen} and its reconstructed image $\hat{x}_{unseen}$, we utilise as baseline the reconstruction error which is defined as the $l2$ norm, *i.e.*, $s_{rec} = \|x_{unseen} - \hat{x}_{unseen}\|_2^2$ between image and reconstructed image (residual).

The second candidate for s is the output of the discriminator. D should give high scores for reconstructions of original, normal images, but low scores for abnormal images, $s_{discr} = 1 - D(x_{unseen})$. Finally, we compute an anomaly score using a gradient-based method, GradCam++, [CSHB18]. Inspired by [KCN+20] ([VPSM19] [LLZ+20]) we apply GradCam++ to the score of the discriminator with regards to the last rectified convolutional layer of the discriminator. This produces attention maps and is also valuable for the localisation of the

pathology. The intuition of using attention maps for computing anomaly scores, is based on the hypothesis that after training the discriminator not only learns to discriminate between normal and abnormal samples but also learns to focus on relevant features in the image [KCN+20]. Thus, specifically for HLHS, where the left artery is missing or is occluded compared to normal samples, a discriminator should identify and locate this difference.

The GradCam++ is computed following:

Let y be the logits of the last layer as they are derived from the discriminator network $D(x_{unseen})$. For the same operators (i, j) and (a, b) applied to the feature map A^k we compute weights:

$$w_k = \sum_i \sum_j \alpha_{ij}^k \text{RELU}(\frac{\partial y}{\partial A_{ij}^k}), \tag{3.6}$$

where the gradient weights a_{ij}^k can be computed as:

$$\alpha_{ij}^k = \frac{\frac{\partial^2 y}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 y}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \{\frac{\partial^3 y}{(\partial A_{ij}^k)^3}\}}, \tag{3.7}$$

and the saliency map (SM) is computed as a linear combination of the forward activation maps followed by a ReLU:

$$SM_{ij} = RELU(\sum_k w_k A_{ij}^k). \tag{3.8}$$

We then computed the sum of the attention maps of image x_{unseen} and its reconstruction from the Generator network, $\hat{x}_{unseen}$:

$$M = \text{SM}(D_{x_{unseen}}) + \text{SM}(D_{\hat{x}_{unseen}}) \tag{3.9}$$

and finally computed the anomaly score s_{attn} as

$$s_{attn} = \frac{\|M \times (x_{unseen} - \hat{x}_{unseen})\|_2^2}{\|M\|_2^2} \tag{3.10}$$

To compute the anomaly score we encapsulate the information of reconstruction [KCN+20]. The model should generalize better, well-reconstructing normal data but failing to reconstruct accurately the anomalous cases. Finally, we attempt to combine anomaly scores, such as s_{rec} with s_{discr}. However, the anomaly detection performance does not improve noteworthily.

3.2.2 Data

The available dataset contains 2D ultrasound images of four-chamber cardiac views. These are standard diagnostic views according to [NHS15]. The images contain labelled examples from normal fetal hearts and hearts with Hypoplastic Left Heart Syndrome (HLHS) [HLH19] from the same clinic, using exclusively an Aplio i800 GI system for both groups to avoid systematic domain differences. HLHS is a birth defect that affects normal blood flow through the heart. It affects a number of structures on the left side of the heart that do not fully develop.

Our dataset consists of 2317 4-chamber view images for which 2224 cases are normal and 93 are abnormal cases. Healthy control view planes have been automatically extracted from examination screen capture videos using a Sononet network [BKM+17a] and manual cleaning from visually trivial classification errors. A set of HLHS view planes that would resemble a 4-chamber view in healthy subjects has been extracted with the same automated Sononet pipeline. Another set has been manually extracted from the examination videos by a fetal cardiologist and 38 cases that are not within 19+0 - 20+6 weeks or show a mix of pathologies have been rejected.

For training, 2131 4-chamber view images (from the available 2224 normal cases), which are considered as normal cases are used. During training, only images from normal fetuses are used. For testing, two different datasets are derived for three different testing scenarios:[3]

For $\mathbf{dataset_1}$(Figure 3.3) we use 4-chamber views from all available HLHS cases, extracted by Sononet and cleaned from gross classification errors; in total 93 cases. Further 93 normal cases have been randomly selected from the remaining test split of the healthy controls and added to this dataset. HLHS cases are challenging for Sononet, which has been trained only on healthy views. Thus, in HLHS cases, it will only select views that are close to the feature distribution of healthy 4-chamber views, which are not necessarily the views a clinician would have chosen. For $\mathbf{dataset_2}$(Figure 3.3), we use the 93 normal cases from $dataset_1$ and the expert-curated HLHS images from the remaining, nonexcluded 53 cases. For each of these cases 1 to 4 different view planes have been identified as clinically conclusive. With this dataset we perform two different

[3]In order to build the test set we have only 93 abnormal cases which are used for testing. In order to choose a number of normal cases, visual inspection was performed and an attempt was made to select cases corresponding to the whole scale of difficulty.

subject-level experiments: a) selecting one of the four frames randomly and b) using all of the 177 clinically selected views in these 53 subjects and fusing the individual abnormality scores to gain a subject-level assessment. We also evaluate per-frame anomaly results. The images

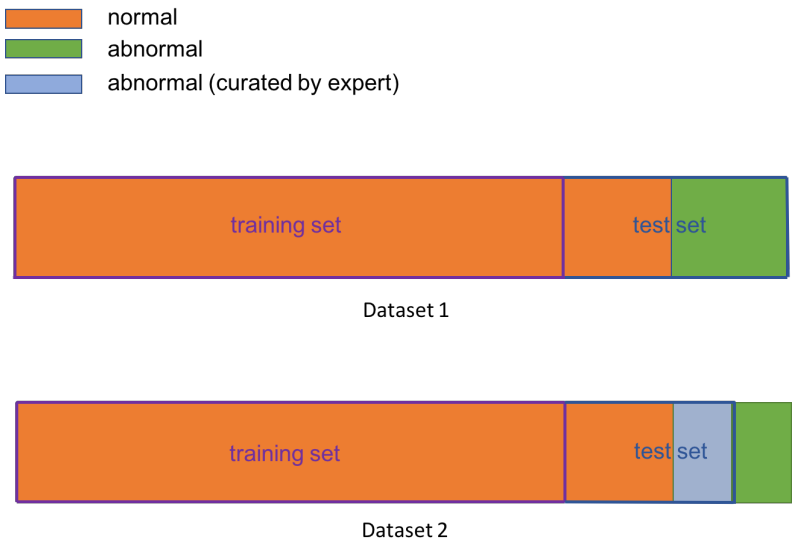


Figure 3.3: Graphical description of Dataset 1 and Dataset 2

are rescaled to 64×64 and normalised to a $[0, 1]$ value range. No image augmentation is used.

3.3 Evaluation and Results

We evaluate our algorithm both quantitatively as well as qualitatively. The capability of the proposed method to localise the pathology is also examined.

3.3.1 Quantitative analysis

For evaluation purposes, the anomaly score is computed as described in Section 3.2.1. For α-GAN and VAE-GAN we use s_{attn}, s_{rec} and s_{discr} as anomaly scores as they are presented in Section 3.2.1. For comparison with the state-of-the-art we train four algorithms: convolu-

tional autoencoder (CAE) [MF15, MMCS11], One-class Deep Support Vector Data Description (DeepSVDD) [RVG+18] and f-AnoGAN [SSW+19], VAE-GAN [BWAN18] [DB16].

Deep Convolutional autoencoder (DCAE) [MF15, MMCS11] is also trained as a baseline. For training, MSE loss is utilised. For DCAE and One-class DeepSVDD we use the same architectures as the ones used for the CIFAR10 dataset in the original work [RVG+18]. Reconstruction error, *i.e.*, $\|x - De(E(x))\|_2$, is defined as anomaly score (s_{DCAE}).

Deep Support Vector Data Description (DeepSVDD) [RVG+18] computes the hypersphere of minimum volume that contains every point in the training set. By minimising the sphere's volume, the chance of including points that do not belong to the target class distribution is minimised. Since in our case all the training data belongs to One-Class (negative class-healthy data) we focus on [RVG+18] . Let $f(.; w)$ be a (deep) neural network with L layers and w^l are the weights of the l_{th} layer. We denote the center of the hypersphere as o. The objective of the network is to minimize the loss which is defined as:

$$\mathcal{L}_{SVDD} = \min_{w} \frac{1}{N} \sum_{i=1}^{N} \|f(x_i; w) - o\|^2 + \frac{\lambda}{2} \sum_{l=1}^{L} \|w^l\|_F^2.$$

The center o is set to be the mean of outputs which is obtained at the initial forward pass. The anomaly score (s_{svdd}) is then defined at inference stage as the distance between a new test sample to the center of the hyper-sphere, *i.e.*, $\|f(x; w_*) - o\|^2$, where w_* are the network parameters of the trained model [RVG+18].

f-AnoGAN [SSW+19] is described in Section 3.1.2. We were not able to successfully train f-AnoGAN using the same networks as we used for α-GAN, hence we utilise similar networks and a similar training framework as described in [SSW+19]. We follow the izi_f training procedure for the encoder network. As anomaly detection score (s_{anogan}) a combination of $L2$ residual loss between the image and its reconstruction and the $L2$ norm of the discriminator's features of an intermediate layer is utilised as it is defined in Table 3.1.

In all algorithms the latent dimension is chosen as 128. We run all experiments 5 times using different (random) initialization seeds [MLKM+18]. We report the average precision, recall

at the Youden index[4] of the Receiver Operating Characteristic (ROC) curves as well as the average corresponding area under curve (AUC) of the 5 runs of each experiment.[5] Furthermore, we apply the DeLong's test [6] [DDCP88] to obtain z-scores and p-values in order to test how statistically different the AUC curve of the proposed model compared to the corresponding curves of the state-of-the-art models (CAE and DeepSVDD, f-AnoGAN and VAE-GAN) is. We perform four different experiments:

Experiment 1 uses $dataset_1$ and aims to evaluate general, frame-level outlier detection performance, including erroneous classifications and fetuses below the expected age range. In Table 3.2, the best performing model based on AUC score is the α-GAN method using s_{attn} as anomaly score which achieves an average of 0.82 ± 0.012 AUC. The α-GAN model achieves the best precision score. However, regarding F1 score and Recall VAE-GAN outperforms α-GAN with 0.78 and 0.88 respectively. DeepSVDD shows the best specificity at 0.76. Figure 3.4 shows the ROC for the best performing (AUC, F1) initialisation and the distribution of normal and abnormal scores for the best model of α-GAN at the Youden index. We present confusion matrices for the α-GAN and the VAE-GAN models in Figure 3.4c and Figure 3.4d. For normal cases both models achieve similar classification performance. However, for identifying abnormal cases α-GAN seems to have an advantage.

Based on the DeLong's test, for Exp. 1, for the average scores (of five experiments), α-GAN compared to f-AnoGAN yields $z = -5.22$ and $p = 1.80e-07$. Similarly, the values for α-GAN compared to CAE are $z = -4.82$ and $p = 1.37e-06$. Finally, comparing α-GAN and DeepSVDD results in $z = -6.49$ and $p = 8.52e-11$. Since $p < 0.01$ for all comparisons, we

[4] Youden index can be defined as $J = \max_c\{\text{Sensitivity}(c) + \text{Specificity}(c) - 1\}$. The cut-point ($c$) that achieves this maximum is referred as the optimal cutoff value (c^*). Youden index is an optimal tradeoff between sensitivity and specificity. It ranges from 0 to 1 [RPWS08, Sha15, RNS18].

[5] In order to evaluate the performance of a machine learning algorithm in medical diagnosis/prediction (e.g binary diagnosis), multiple factors should be assessed both by machine learning scientists as well as by clinical experts [PH18]. Validation of an algorithm is mainly assessed by model's discrimination (i.e binary normal/abnormal) performance (in the test set). The ROC curve is an effective way to determine the discrimination performance of a model. It represents the trade-off between sensitivity and specificity and it is not affected by disease prevalence [Obu03, Flo08]. Also, several measures for the accuracy of the algorithm based on the ROC can be computed (e.g area under curve AUC) [Obu03]. AUC value varies between 0 and 1. An excellent model has AUC score close to 1 (good discrimination performance). Furthermore, it is also important to report a combination of multiple metrics (which also could be intuitive for clinicians) such as the false positives rate and false negative rate [AST+21]. The interpretation of these metrics should be made very carefully based also on the specific domain application. Furthermore, apart from commonly used performance metrics, it would be highly valuable, to verify the model's performance taking into consideration the ultimate patient benefit outcome [PH18].

[6] In order to compare the difference between two AUCs, we apply the *DeLong test*. This method provides a confidence interval and standard error of the difference between two correlated AUCs [DPD12]. DeLong's test can be used to compare the area under two or more correlated ROC curves [SX14].

can assume that α-GAN performs significantly better than the state-of-the-art when applied to fetal cardiac ultrasound screening for HLHS. Comparing α-GAN with VAE-GAN the values are $z = -1.21$ and $p = 0.22$ which does not indicate a significant difference between AUC curves. As can be seen from the results, the GAN-based methods achieve better performance for detecting HLHS.

Quantitative performance scores					
Method	Precision	Recall	Specificity	F1 score	AUC
CAE [RVG+18]	0.65 ± 0.027	0.64 ± 0.061	0.65 ± 0.074	0.64 ± 0.061	0.65 ± 0.016
DeepSVDD [RVG+18]	0.67 ± 0.106	0.37 ± 0.258	$\mathbf{0.76} \pm 0.260$	0.41 ± 0.150	0.53 ± 0.039
f-AnoGAN [SSW+19]	0.58 ± 0.022	0.58 ± 0.130	0.59 ± 0.097	0.57 ± 0.072	0.57 ± 0.039
s_{rec} (VAE-GAN)	0.69 ± 0.018	$\mathbf{0.88} \pm 0.060$	0.61 ± 0.057	$\mathbf{0.78} \pm 0.015$	0.78 ± 0.010
s_{discr} (VAE-GAN)	0.75 ± 0.220	0.29 ± 0.360	0.75 ± 0.360	0.27 ± 0.230	0.42 ± 0.027
s_{attn} (VAE-GAN)	0.72 ± 0.014	0.83 ± 0.043	0.68 ± 0.037	0.77 ± 0.014	0.79 ± 0.008
s_{rec} (α-GAN)	0.64 ± 0.017	0.87 ± 0.054	0.50 ± 0.038	0.74 ± 0.024	0.71 ± 0.029
s_{discr} (α-GAN)	0.65 ± 0.056	0.51 ± 0.240	0.70 ± 0.205	0.53 ± 0.170	0.61 ± 0.067
s_{attn} (α-GAN)	$\mathbf{0.73} \pm 0.026$	0.82 ± 0.068	0.70 ± 0.059	0.77 ± 0.029	$\mathbf{0.82} \pm 0.012$

Table 3.2: Anomaly detection performance for Exp. 1 using $dataset_1$. Best performance in bold.

Experiment 2 uses $dataset_2$ for specific disease detection capabilities with expert-curated, clinically conclusive 4-chamber views for 53 HLHS cases. We choose one of the relevant views per subject randomly. Table 3.3 summarises these results. VAE-GAN has the highest AUC, F1, precision, recall and specificity scores using s_{attn} as anomaly score. Also, we note from Figure 3.5c and Figure 3.5d that the VAE-GAN method misclassified less HLHS cases while achieving better performance for confirming normal cases. Average F1 score is 0.89. Figure 3.5 shows ROC, anomaly score distribution and confusion matrices at the Youden index of this experiment.

Quantitative performance scores					
Method	Precision	Recall	Specificity	F1 score	AUC
CAE [RVG+18]	0.63 ± 0.095	0.56 ± 0.120	0.78 ± 0.130	0.57 ± 0.025	0.72 ± 0.015
DeepSVDD [RVG+18]	0.39 ± 0.016	0.80 ± 0.160	0.28 ± 0.160	0.52 ± 0.032	0.49 ± 0.038
f-AnoGAN [SSW+19]	0.56 ± 0.077	0.52 ± 0.097	0.75 ± 0.140	0.53 ± 0.041	0.63 ± 0.043
s_{rec} (VAE-GAN)	0.64 ± 0.067	0.80 ± 0.060	0.74 ± 0.078	0.71 ± 0.020	0.84 ± 0.009
s_{discr} (VAE-GAN)	0.36 ± 0.220	0.56 ± 0.450	0.46 ± 0.430	0.34 ± 0.205	0.39 ± 0.037
s_{attn} (VAE-GAN)	$\mathbf{0.71} \pm 0.046$	$\mathbf{0.85} \pm 0.038$	$\mathbf{0.80} \pm 0.058$	$\mathbf{0.77} \pm 0.016$	$\mathbf{0.89} \pm 0.009$
s_{rec}(α-GAN)	0.59 ± 0.050	0.81 ± 0.060	0.66 ± 0.010	0.68 ± 0.015	0.79 ± 0.030
s_{discr}(α-GAN)	0.48 ± 0.100	0.51 ± 0.280	0.61 ± 0.280	0.43 ± 0.110	0.53 ± 0.030
s_{attn}(α-GAN)	0.59 ± 0.098	0.76 ± 0.150	0.66 ± 0.180	0.64 ± 0.037	0.77 ± 0.046

Table 3.3: Anomaly detection performance using $dataset_2$ for Exp. 2. Best performance in bold.

Experiment 3 uses $dataset_2$ and is similar to Exp. 2 except that we take all clinically identified views for each subject into account. We average the individual anomaly scores for each frame, depending on the number of frames that are available per subject. VAE-GAN achieves a better AUC score with 0.86 compared to 0.84 of α-GAN as can be seen in Table 3.4. However, as can

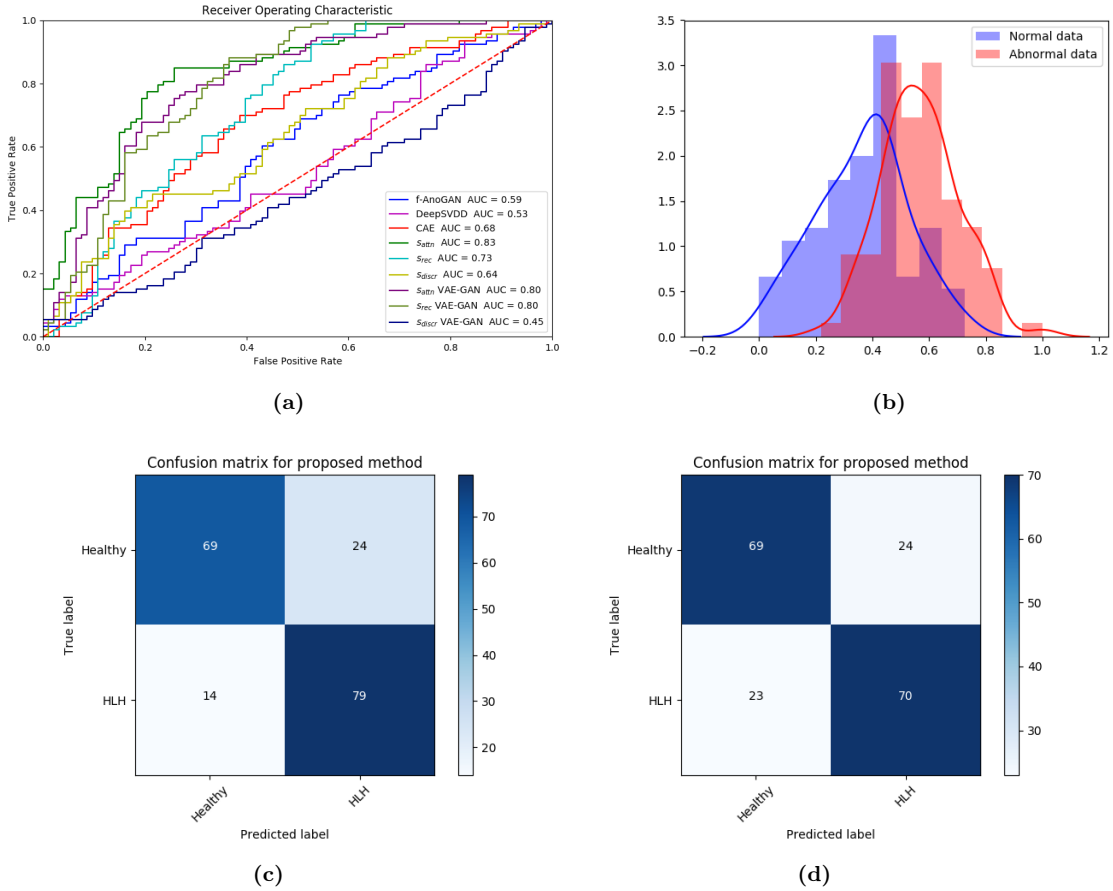


Figure 3.4: (a) ROC-AUC curves in Exp. 1; (b) Distribution of normal/abnormal score values for the α-GAN model with s_{attn} as anomaly score (c) Confusion matrix for the best performing run of the proposed α-GAN (d) Confusion matrix for the best performing run of the VAE-GAN. This figure focuses on the results of the best performing initialisation from five experiments with α-GAN (or VAE-GAN) while Table 3.2 shows average metrics.

be seen from the confusion matrices (best performing initialisation), α-GAN shows a better true positive rate at the cost of a higher number of false positives (Figure 3.6c). This configuration might be preferred in a clinical setting since it reduces the number of missed cases at the cost of a slightly higher number of false referrals.

Experiment 4 is similar with the Exp. 3 except that we evaluate frame-level performance in Table 3.5. VAE-GAN is again better in terms of precision and AUC performance. However, similar to Exp. 3 α-GAN has an advantage when recognising the cases with pathology at a cost of a higher false positive rate.

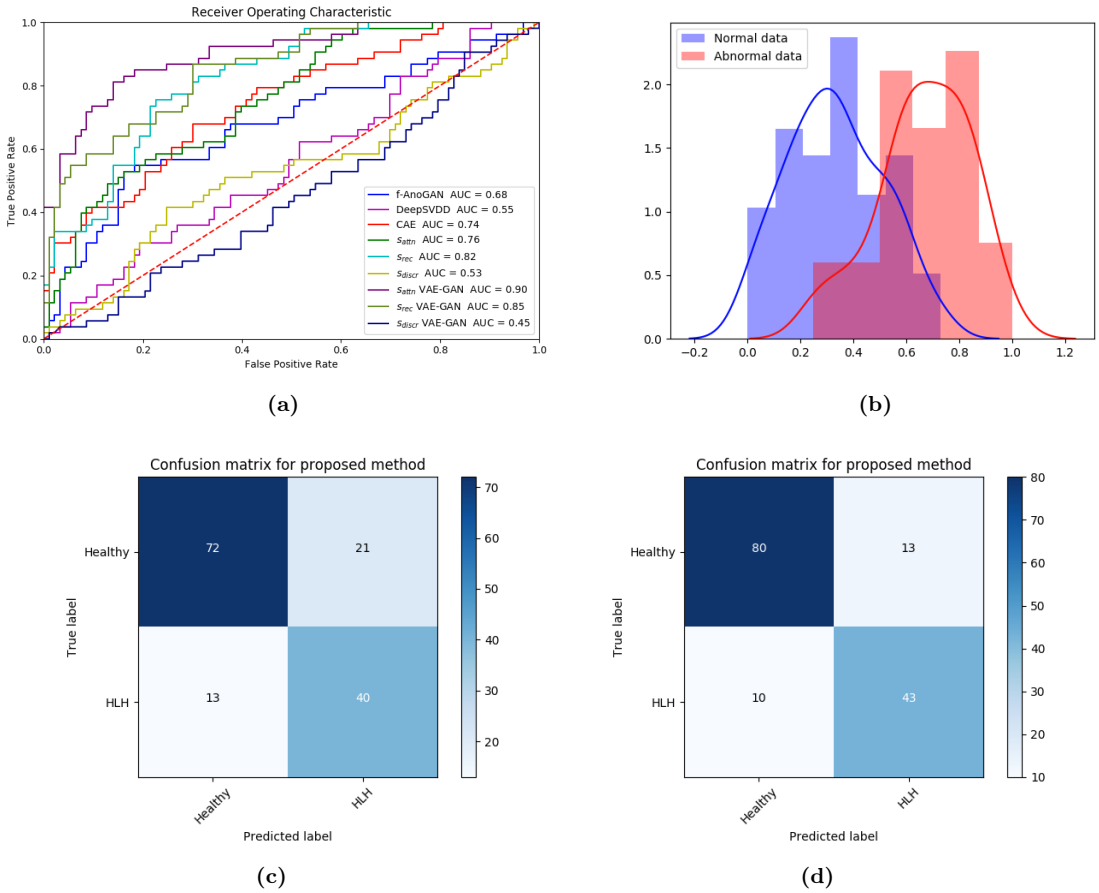


Figure 3.5: *dataset*$_2$, Exp. 2: (a) ROC-AUC curves in Exp. 2; (b) Distribution of normal/abnormal score values for the VAE-GAN model with s_{attn} as anomaly score (c) Confusion matrix for the best performing run using s_{rec} of the proposed α-GAN. (d) Confusion matrix for the best performing run using s_{attn} of the VAE-GAN.

Quantitative performance scores					
Method	Precision	Recall	Specificity	F1 score	AUC
CAE [RVG+18]	0.51 ± 0.061	0.80 ± 0.136	0.54 ± 0.150	0.61 ± 0.018	0.70 ± 0.024
DeepSVDD [RVG+18]	0.42 ± 0.063	0.69 ± 0.312	0.39 ± 0.311	0.47 ± 0.140	0.48 ± 0.038
f-AnoGAN [SSW+19]	0.55 ± 0.029	0.79 ± 0.067	0.62 ± 0.068	0.64 ± 0.016	0.74 ± 0.013
s_{rec} (VAE-GAN)	0.60 ± 0.029	0.87 ± 0.049	0.67 ± 0.056	0.71 ± 0.014	0.81 ± 0.076
s_{discr} (VAE-GAN)	0.37 ± 0.150	0.98 ± 0.400	0.032 ± 0.39	0.53 ± 0.021	0.14 ± 0.034
s_{attn} (VAE-GAN)	$\mathbf{0.66} \pm 0.036$	0.88 ± 0.035	$\mathbf{0.74} \pm 0.050$	$\mathbf{0.75} \pm 0.014$	$\mathbf{0.86} \pm 0.017$
s_{rec} (α-GAN)	0.57 ± 0.041	0.86 ± 0.091	0.62 ± 0.098	0.68 ± 0.022	0.78 ± 0.019
s_{discr} (α-GAN)	0.42 ± 0.035	0.89 ± 0.110	0.28 ± 0.155	0.57 ± 0.067	0.48 ± 0.017
s_{attn} (α-GAN)	0.62 ± 0.040	$\mathbf{0.92} \pm 0.100$	0.67 ± 0.069	0.73 ± 0.024	0.84 ± 0.018

Table 3.4: Anomaly detection performance on subject level for *dataset*$_2$ and Exp. 3. Best performance in bold.

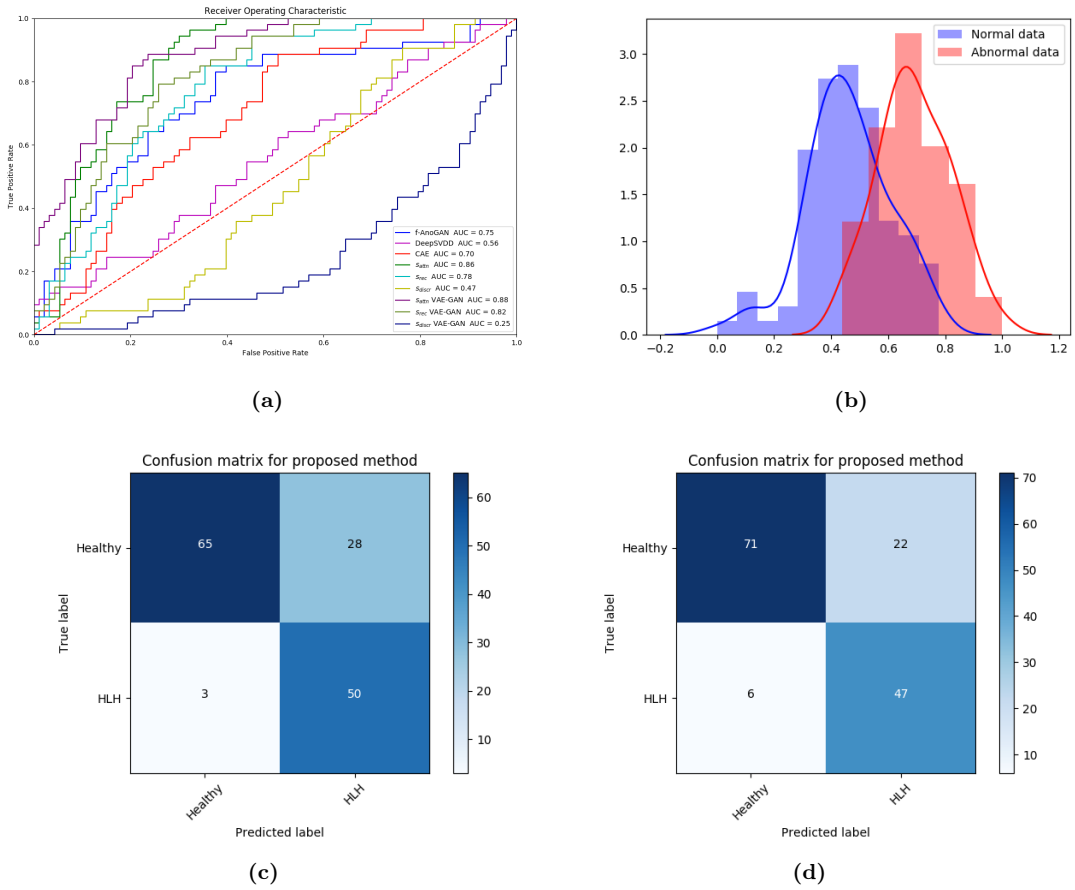


Figure 3.6: *dataset*$_2$, Exp. 3: (a) ROC-AUC curves in Exp. 3; (b) Distribution of normal/abnormal score values for the VAE-GAN model with s_{attn} as anomaly score (c) Confusion matrix for the best performing run of the proposed α-GAN (d) Confusion matrix for the best performing run of the VAE-GAN.

Quantitative performance scores					
Method	Precision	Recall	Specificity	F1 score	AUC
CAE [RVG+18]	0.80 ± 0.026	0.57 ± 0.081	0.71 ± 0.075	0.66 ± 0.051	0.67 ± 0.020
DeepSVDD [RVG+18]	0.86 ± 0.100	0.09 ± 0.030	$\mathbf{0.96} \pm 0.025$	0.15 ± 0.053	0.44 ± 0.025
f-AnoGAN [SSW+19]	0.82 ± 0.041	0.56 ± 0.070	0.75 ± 0.095	0.66 ± 0.040	0.66 ± 0.013
s_{rec} (VAE-GAN)	0.82 ± 0.023	0.74 ± 0.062	0.69 ± 0.073	0.77 ± 0.024	0.77 ± 0.009
s_{discr} (VAE-GAN)	0.80 ± 0.130	0.03 ± 0.007	0.99 ± 0.008	0.05 ± 0.012	0.37 ± 0.047
s_{attn} (VAE-GAN)	$\mathbf{0.86} \pm 0.016$	0.78 ± 0.051	0.76 ± 0.046	0.82 ± 0.023	$\mathbf{0.82} \pm 0.023$
s_{rec} (α-GAN)	0.80 ± 0.016	0.80 ± 0.032	0.62 ± 0.051	0.80 ± 0.012	0.75 ± 0.017
s_{discr} (α-GAN)	0.71 ± 0.060	0.72 ± 0.300	0.38 ± 0.320	0.66 ± 0.180	0.48 ± 0.055
s_{attn} (α-GAN)	0.82 ± 0.030	$\mathbf{0.85} \pm 0.110$	0.64 ± 0.094	$\mathbf{0.83} \pm 0.047$	0.81 ± 0.018

Table 3.5: Anomaly detection performance using *dataset*$_2$ in Exp. 4 for evaluation per frame. Best performance in bold.

3.3.2 Qualitative analysis

In order to evaluate the ability of the algorithm to localise anomalies, we plot the class activation maps as they are derived from the proposed model. In Figure 3.8, we present(overlay) the

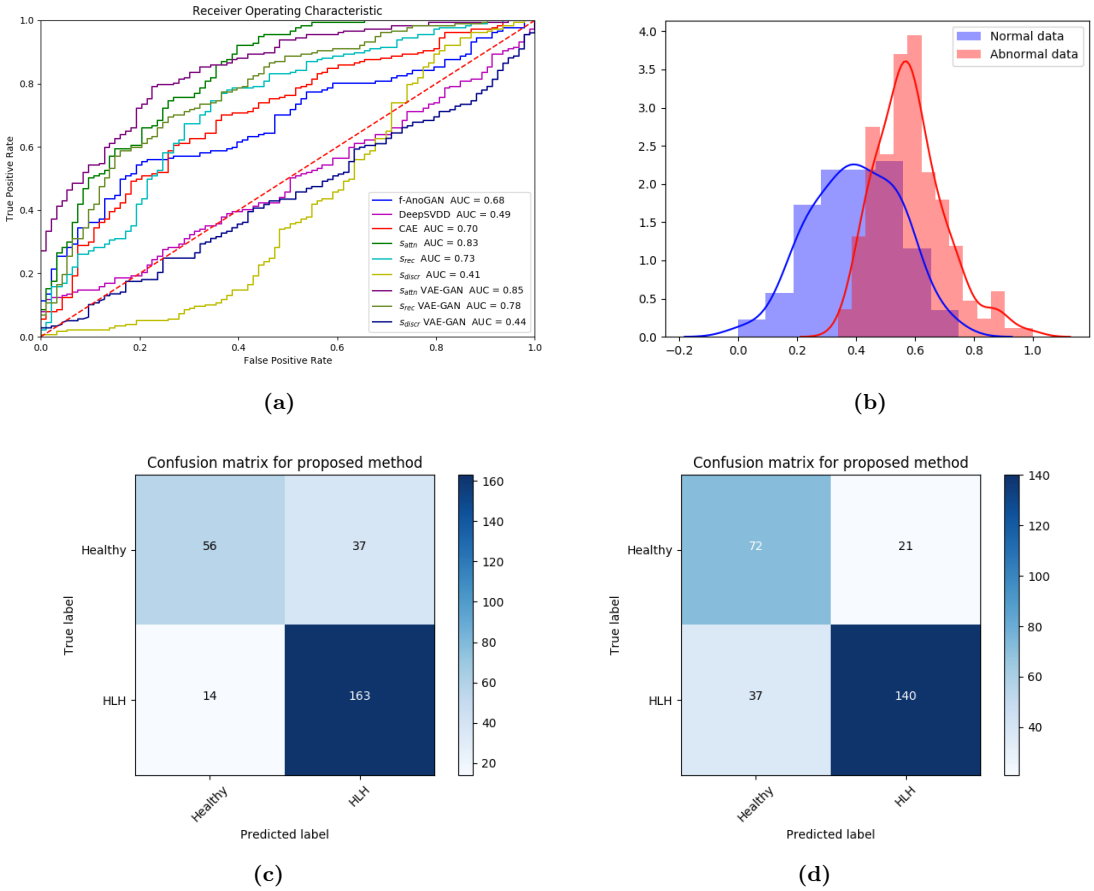


Figure 3.7: *dataset*$_2$, Exp. 4: (a) ROC-AUC curves in Exp. 4; (b) Distribution of normal/abnormal score values for the α-GAN model with s_{attn} as anomaly score (c) Confusion matrix for the best performing run of the proposed α-GAN. (d) Confusion matrix for the best performing run of the VAE-GAN. This figure focuses on the results of the best performing initialisation from five experiments with α-GAN (or VAE-GAN) while Table 3.2 shows average metrics.

obtained saliency maps from the GradCam++(Eq. 3.8), onto pathological images in *dataset*$_1$ (Exp.1). In the abnormal cases, attention focuses exactly on the area of heart. As a consequence, anomaly scores in such cases are higher compared to normal cases and they are correctly indicated as anomalous. All anomaly scores are normalised in the range of $[0, 1]$. There are cases that our algorithm fails to classify correctly. Either they are abnormal and they are classified as normal (False Negative-FN) or they are healthy and identified as anomalous (False Positive-FP). In Figure 3.9 examples for False Positive cases are presented alongside False Negative cases. Bad image reconstruction quality is a limiting factor for deriving a representative anomaly score for a data sample (the anomaly score contains a reconstruction component). For

instance, in some reconstructions either a part of the heart is missing (left or right ventricle/ atrium) or the shape of the heart is quite different from a normal heart (*e.g.*, a very "long" ventricle). As a consequence, not only the reconstruction error is high, but also the attention mechanism focuses in this area, since it is recognised (by the network) as anomalous. Consequently, the total anomaly score is high. In fewer examples the signal-to-noise ratio (SNR) is low, *i.e.*, images are blurry, and so the network fails to reconstruct the images at all. Furthermore, in the False Positive examples Figure 3.9a, from clinical perspective, the angle is not quite right, so it makes the ventricles look shorter than they are. This confuses the model, forcing the discriminator's attention to indicate this area as anomalous. Another point which is very interesting to highlight, is that there are cases where some frames are very difficult, even for experts. Such an example is given in Figure 3.9b, where although the second image from left belongs to an abnormal subject, the specific frame appears normal at the first glance. Such cases also highlight limitations of single-view approaches. In practice, all relevant frames showing the 4ch view could be processed with our method and a majority vote could regarding referral be calibrated on a ROC curve.

All the above plots and comparisons utilise the top-1 performing experiment among all the runs of the experiments for α-GAN.

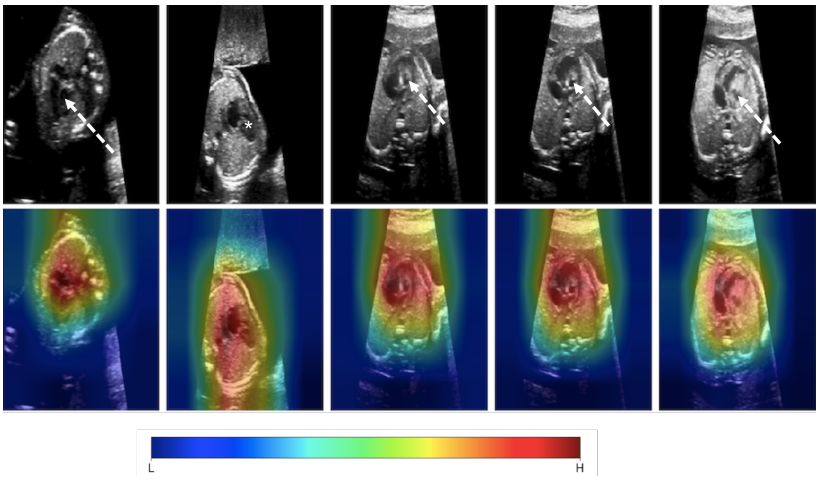


Figure 3.8: Top row: Pathological subjects Bottom row: GradCam++ visualisation of attention maps using α-GAN (Exp. 1).
*= dominant RV with no visible LV cavity, solid white arrow = deceptively normal-looking LV, dashed white arrow = globular, hypoplastic LV

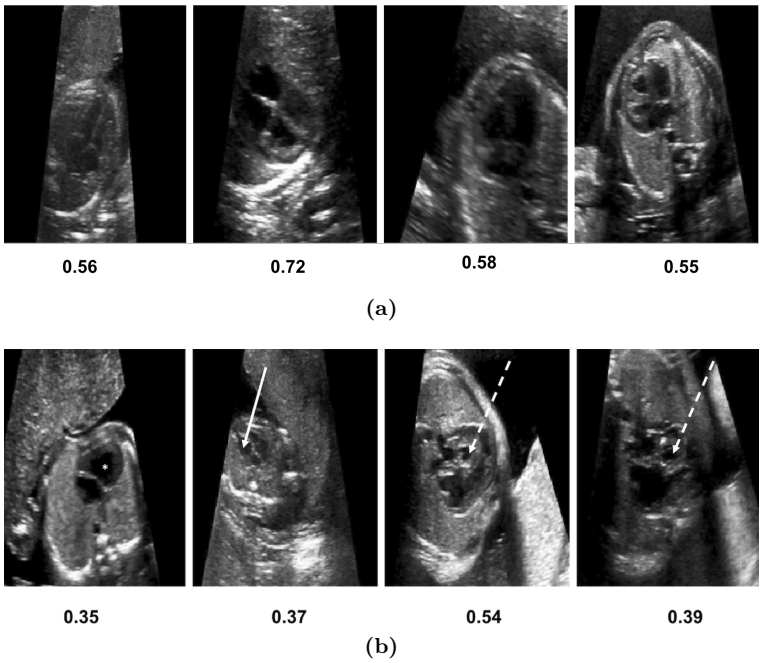


Figure 3.9: (a) Examples of False Positive along with the anomaly scores s_{attn} (b) False Negative cases along with the anomaly scores s_{attn} (Exp. 1). *= dominant RV with no visible LV cavity, solid white arrow = deceptively normal-looking LV, dashed white arrow = globular, hypoplastic LV. Low Signal-to-Noise Ratio (SNR)

3.4 Discussion

Our results are promising and confirm that automated anomaly detection can work in fetal 2D ultrasound as shown on the example of HLHS. For this pathology we achieve an average accuracy of 0.81 AUC, improving significantly the detection rate of front-line-of-care sonographers during screening, which is often below 60% [CHRP07]. However, there are open issues.

False negative rates are critical for clinical diagnosis and downstream treatment. In a clinical setting, a method with zero false negative predictions would be preferred, *i.e.*, a method that *never* misses an anomaly, but potentially predicts a few false positives. Assuming that the false positive rate of such an algorithm is significantly below the status quo, the benefits for antenatal detection and potentially better postnatal outcomes would outweigh the costs.

Of course, an algorithm with a 100% false positive rate is also not desirable, hence calibration on the ROC must be performed.

A key aspect of the proposed algorithm is the ability of the discriminator to highlight decisive

areas in images. In order to achieve this, it is necessary to produce good reconstructions of normal images. However, reconstruction quality can be limited, depending on the given sample. A larger dataset could provide a mitigation strategy for this. Furthermore, alternative ways for visualising attention could be explored for disease-specific applications such as implicit mechanisms of attention like attention gates [SOS+19].

Although we have experimented with different type of noise (e.g Uniform) and various augmentation techniques (e.g horizontal flip) we did not notice an improvement in anomaly detection performance. However, a further investigation of other augmentation techniques should be done. For instance, elastic transformation or grid distortions could be utilised in the training process. Furthermore, it could be useful to examine the impact of (non-linear) mixed-example data augmentation [LGN21], or physics-inspired transformations [TES+21] which has been applied specifically to ultrasound images (i.e domain-specific data augmentation). Finally, another interesting approach is to test whether generative adversarial networks-based augmentation methods could be utilised efficiently to ultrasound images.

Moreover, it would be interesting to explore the sensitivity of our method for other sub-types of congenital heart disease. Intuitively, accuracy of a general anomaly detection method should be similarly high for other syndromes that affect the morphological appearance of the fetal four-chamber view. HLHS has a particularly grossly abnormal appearance. There are a lot of other CHD examples with a subtly abnormal 4ch view that would probably be much harder to detect even for human experts. Additionally, in practice, confounding factors may bias anomaly detection methods towards more obvious outliers, while subtle signs of disease or indicators encoded in other dimensions like the spatio-temporal domain may still be missed.

Finally, robust time-series analysis is still a challenging fundamental research question and we are looking forward to extending our method to full video sequences in future work. For instance, adding a Long Short-Term Memory Network (LSTM)-based component in our encoder-decoder could prove beneficial [MS16] for this purpose. In this way, we could incorporate, model and learn the spatio-temporal representations of a video sequence.

3.5 Conclusion

In this Chapter, we attempt to consider the detection of congenital heart disease as a one-class anomaly detection problem, learning only from normal samples.

The proposed unsupervised architecture shows promising results and achieves better performance compared to existing state-of-the-art image anomaly detection methods. However, since clinical practice requires highly reliable anomaly detection methods, further work will need to be done, in order to deal with the aforementioned weaknesses. In this way, we could avoid false positives in order to mitigate patient stress and strain on healthcare systems and false negatives to prevent missed diagnoses.

Chapter 4

Exploring the Relationship Between Segmentation Uncertainty, Segmentation Performance and Inter-rater Variability with Probabilistic Networks

Medical image segmentation is an essential tool for clinical decision making and treatment planning. Automation of this process led to significant improvements in diagnostics and patient care, especially after recent breakthroughs that have been triggered by deep learning. However, when integrating automatic tools into patient care, it is crucial to understand their limitations and to have means to assess their confidence for individual cases. Uncertainty quantification has been the subject of recent research. Methods have been developed to calculate the segmentation uncertainty automatically during the inference stage. Varying image quality and different levels of human annotator expertise are an integral part of aleatoric uncertainty. However, it is unknown how much this variability affects the final segmentation uncertainty. In this Chapter we explore potential links between deep network segmentation uncertainties with inter-observer variance and segmentation performance. In order to evaluate the proposed framework, we utilised the LIDC-IDRI dataset [AIMB+11], which contains multiple expert annotations for each subject.

4.1 Introduction

Segmentation, *i.e.*, delineation of anatomical structures in 2D/3D, is a core necessity in medical image analysis. In most cases, segmentation is carried out manually by an expert. It is well known that manual segmentation suffers from inter-observer variability and that segmentation quality is influenced by factors such as fatigue, different domain knowledge, level of expertise, and image resolution. As a result, manual segmentations contain data uncertainty and can thus be ambiguous for diagnosis or confusing for supervised learning methods. Nevertheless, annotator confidence can be an important source of information for clinical decision making. Varying annotator confidence can be a trigger for additional imaging tests and an indicator for quality control and treatment options. Confidence is an important factor to weigh individual test result but it is only qualitatively assessed in the clinical practice.

In this Chapter we explore whether human inter-observer variability can be correlated with the distribution of two different probabilistic neural networks and investigate the impact of this variability on the estimation of segmentation uncertainty and segmentation performance.

4.1.1 Uncertainty and Inter-rater variability estimation in Deep Neural Networks

Recent successes of deep learning (e.g CNNs, U-Net, LSTM) for image segmentation [RFB15, ÇAL+16, KLN+17] promise to reduce clinical annotation workload. Currently, the majority of these methods lack the ability to communicate annotator confidence. However, quantitative assessment of uncertainties is key to guarantee quality of care, increases trust and can have great impact on therapeutic decisions.

Estimation of uncertainty in medical imaging segmentation has been attempted in works such as [RCNW18, NPAA18, MNM+19]. A reference to the most relevant uncertainty quantification works is made below. A detailed review of uncertainty estimation methods in deep neural networks, with applications in medical image segmentation, was given in Section 2.2.5.

In [RCNW18] authors use Monte Carlo samples from the posterior distribution of a Bayesian fully Convolutional neural network which are derived using dropout at test phase. Based on these samples, they compute structure-wise and voxel-wise uncertainties metrics, which as they

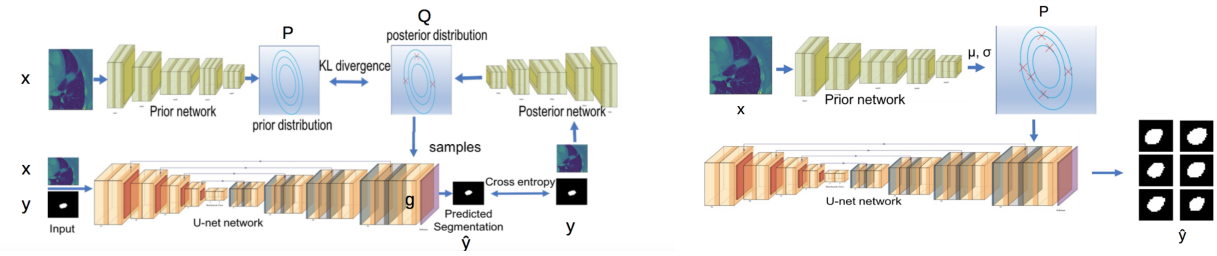
prove, are highly correlated with segmentation accuracy. Application field is brain segmentation. In another work [NPAA18] Monte Carlo dropout is used for uncertainty estimation in Multiple Sclerosis lesion detection and segmentation. Four different voxel-wise uncertainties were utilised including prediction variance, Monte Carlo sample variance, predictive Entropy and Mutual Information. As it was proved by the results, filtering based on uncertainty leads to improvement on the lesion detection accuracy. In [MNM+19] authors propose a framework to approximate Bayesian inference in deep neural networks by imposing Bernoulli distribution directly on the weights of the deep model. Then Monte Carlo samples from posterior distribution are utilised to compute Mutual Information as metric for uncertainty in CT-organ semantic segmentation.

Furthermore, the effect of inter-observer variability for estimation of uncertainty in segmentation is studied in [JME+18]. Authors, in MRI images from brain tumors, explore the impact of different label fusion techniques (e.g no fusion, STAPLE [WZW04], union, intersection, majority) in estimation of segmentation uncertainty. As it is proved, there is a link between uncertainty estimation and inter-observer variability. Monte Carlo dropout is also used in this work for estimation of uncertainty (entropy).

A recent work examines also the effect of inter-rater variability on uncertainty estimation. In [JJJ+19], authors examine the impact that label fusion/sampling techniques and uncertainty estimation (using Ensembles, test-time augmentation, MC-dropout and Monte Carlo Batch Normalization) methods have on model calibration in the skin image classification problem. Finally, an alternative way to produce plausible segmentation hypotheses is proposed in [KRPM+18] where authors use generative segmentation model, a combination of U-Net and conditional variational autoencoder, in order to produce plausible segmentation hypotheses (diverse samples) for lung abnormalities segmentation task [KRPM+18, BKCTC+19].

4.2 Materials and Methods

Two different probabilistic networks are utilised in our work: a $3D$ probabilistic U-Net (**PUNet**) and a $3D$ U-Net using Monte Carlo Dropout during inference (**DUNet**). 3D lung CT scans have been utilised for the experiments in this Chapter. In order to model variability and uncertainty



(a) Training **(b)** Inference

Figure 4.1: PUNet [KRPM+18] as we use it for our method.

efficiently, it would be beneficial, if deep learning models could generate diverse segmentation hypotheses, capture rare segmentation variants and at the same time scale efficiently with the number of generated samples. The PUNet seems to meet these requirements more efficiently compared to deep ensembles [LPB17] or deep networks with M heads [IÇG+18]. On the other hand, DUNet is a commonly used approach for uncertainty estimation based on the multiple MC samples that can be generated at test time using MC-Dropout. Furthermore, both models are U-Net-based architectures and U-Net has shown outstanding performance [RFB15] in segmentation tasks in medical image analysis.

PUNet: We extend a $2D$ probabilistic U-Net [KRPM+18], which is a combination of a U-Net [RFB15, ÇAL+16] and a conditional variational autoencoder [SLY15] to 3D. The whole architecture consists of three networks, which is shown in Figure 4.1.

Let x be an input volume, M the segmentation map, $\hat{y}$ the predicted segmentation, y the ground truth segmentation as it is produced by several experts ($n = 4$ for LIDC), C the number of classes and N number of voxels per volume similar as proposed by [KRPM+18]. The Prior net is conditioned on the input volume x. It computes the distribution over the (low-dimensional) latent space R^K. At inference stage samples that are produced by this distribution are concatenated with the last layer's feature maps of the segmentation network, which produces a segmentation map for each sample. More precisely the prior probability distribution P is modelled as an axis-aligned Gaussian distribution with mean $\mu_{prior}(x; w_{prior}) \in R^K$ and variance $\sigma_{prior}(x; w_{prior}) \in R^K$. To sample T segmentations we apply the network T times to the same input volume. In each iteration a sample $z_t, t = \{1, 2, ..., T\}$ is drawn from the distribution:

$$z_t \sim P(.|x) = \mathcal{N}(\mu_{prior}(x; w_{prior}), diag(\sigma_{prior}(x; w_{prior}))) \tag{4.1}$$

Each sample is reshaped to a K-channel feature map with the same shape as the segmentation map. This feature map is concatenated to the last activation map of a U-Net. Then, a segmentation map, which corresponds to sample z_t, is produced by $M_t = f(g(x, w), z_t, \psi)$ where w is the UNet parameters and ψ weights of the last layer of U-Net.

The final segmentation maps are based on the latent variable z. By sampling from the prior distribution $P(z|x)$ (Eq. 4.1), we can obtain multiple segmentation variants. The posterior net is conditioned on the volume x as well as the ground truth y. It learns to recognize (embeds) segmentation variants $\mu_{post}(x, y; \nu) \in R^K$ with some uncertainty $\sigma_{post}(x, y; \nu) \in R^K$ in the low dimensional latent space. The output is denoted as posterior distribution Q. A sample z from this distribution

$$z \sim Q(.|x, y) = \mathcal{N}(\mu_{post}(x, y; \nu), diag(\sigma_{post}(x, y; \nu)) \tag{4.2}$$

combined with the activation map of the U-Net will result in a predicted segmentation $\hat{y}$. The loss function is composed by two terms. The first is the cross entropy loss

$$E_{z\sim Q(.|y,x)}[-\log P_c(y|M(x, z))]$$

which penalizes the difference between the ground truth and the segmentation map. The second one is the Kullback-Leibler (KL) divergence $D_{KL}(Q(z|y, x)||P(z|x))$ which penalizes differences between the posterior distribution Q and the prior distribution P. Both terms are combined as a weighted sum with a weighting factor β as proposed by [KRPM+18]. Thus, the total loss function is defined as:

$$L(y, x) = E_{z\sim Q(.|y,x)}[-\log P_c(y|M(x, z))] + \beta * D_{KL}(Q(z|y, x)||P(z|x)) \tag{4.3}$$

In our experiments we use $\beta = 0.2$. Differences between training and inference are outlined in Figure 4.1.

DUNet: We utilise a U-Net where dropout layers are activated during inference. Dropout can be explained as an alternative way to perform Variational inference [GG16b]. Cross entropy between ground truth and predicted segmentation is utilised as loss function.

To produce multiple diverse segmentation hypotheses, we utilise PUNet and DUNet. In order

to exploit volumetric information, $3D$ versions of the above models are trained using $3D$ convolutions. The U-Nets consist of 3 layers. Each layer consists of $3D$ convolution blocks followed by Rectified Linear Unit (ReLU) activation, batch normalization and max pooling. Filter size is $3 \times 3 \times 3$. We start the number of feature maps at 32 and double it after each block. For the prior net as well as for the posterior net in the PUNet, we utilize the encoder part of the U-Net. We train the networks using exponential decay learning rate and the Adam optimizer. For the DUnet, dropout is used after each layer in the encoding part of U-Net. We use a dropout probability of 0.2. We generate an equal number of samples T for both $3D$ networks. All networks are implemented in Python using Tensorflow, on a workstation with NVIDIA Titan X GPU. In order to estimate uncertainty we compute two uncertainty scores: Z_{var} and Z_S using variance [KBC17, SG18] and predictive entropy [GIG17] of samples respectively. We define mean variance across all classes C as:

$$\sigma^2(x^*) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}(p_t(y=c|x^*,w) - \hat{p}(y=c|x^*,w))^2, \tag{4.4}$$

where $\hat{p}(y|x^*,w)$ is the average of softmax probabilities of T samples for each $c \in [1,...,C]$ and p_t the output of the network for sample t. Subsequently we define Z_{var} as

$$Z_{var} = \frac{1}{N}\sum_{v=1}^{N}\sigma^2(x^*(v)), \tag{4.5}$$

and predictive entropy S as

$$S(x^*) = -\sum_{c=1}^{C}\hat{p}(y=c|x^*,w) \times \log(\hat{p}(y=c|x^*,w)). \tag{4.6}$$

Thus, for each subject x^*, Z_S is computed as:

$$Z_S = \frac{1}{N}\sum_{v=1}^{N}S(x^*(v)) \tag{4.7}$$

We utilise the Sørensen–Dice coefficient (Dice score) to characterise segmentation performance. To examine possible linear correlation between segmentation performance and segmentation uncertainty, we compute the Pearson correlation coefficient (ρ) between Z_S and Z_{var} and the Dice score.

To investigate the relationship between Z_S and Z_{var} and the variability among human experts we define the area of human disagreement (Γ) using a process similar to XOR ($\oplus$) (for two raters) of the different annotations for each subject. For each voxel the $\oplus$ operation will result 1 indicating that at least one annotator disagrees (disagreement) while 0 is used where all annotators agree (agreement). However, for four annotators, the process is slightly modified and the output is computed (for all possible combinations) as it is presented in the Figure 4.2. For a fair comparison with Z_S and Z_{var} we utilize the same schemes for deriving quantitative

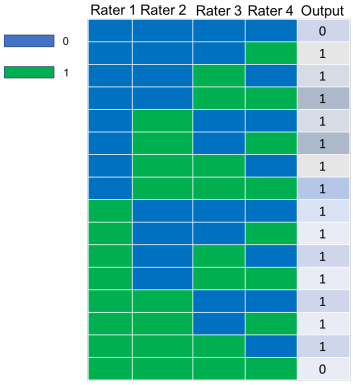


Figure 4.2: Description of the process of defining the human disagreement area

uncertainty: predictive Entropy(S) Eq. 4.6 and variance σ^2 Eq. 4.4. For qualitative analysis we emply Mutual Information (MI) [SG18, MNM+19] and a map of Softmax output probabilities for the predominant class (Softmax). MI is defined using Eq. 4.6 [SG18, Gal16], *i.e.*,

$$MI(x^*) = S(x^*) + \frac{1}{T}\sum_{c,t} p_t(y = c|x^*, w)\log(p_t(y = c|x^*, w)), \tag{4.8}$$

where $p_t(y = c|x^*, w)$ is the the softmax output of the network for each sample. We then characterise a voxel v of a new sample x^* as certain/uncertain using

$$x^*(v) = \begin{cases} \text{uncertain, if } S(x^*(v)) >= \theta, \\ \text{certain, otherwise,} \end{cases} \tag{4.9}$$

where θ is a threshold and v a voxel. Alternatively, we can replace $S(x^*)$ in Eq. 4.9 with variance σ^2 for estimation of uncertainty. We use a threshold since we assume that human perception of uncertainty is more accurate when interpreted binary than continuous. Evidence for this is given in behavioural sciences literature, *e.g.* [FR03, HP02].

We compute the ROC curve between True Positive Rate (TPR) and False Positive Rate (FPR) for the binary case. This allows us to correlate segmentation uncertainty with expert uncertainty. With Γ and Eq. 4.9 we define TPR and FPR as

$$\text{TPR} = p(\text{uncertain}|\text{disagreement}) = \frac{p(\text{uncertain}, \text{disagreement})}{p(\text{disagreement})}, \tag{4.10}$$

and

$$\text{FPR} = p(\text{uncertain}|\text{agreement}) = \frac{p(\text{uncertain}, \text{agreement})}{p(\text{agreement})}. \tag{4.11}$$

To evaluate segmentation uncertainty with respect to Γ we use disagreement accuracy (*DisAcc*) as metric [MNM+19, SDVG19]. *DisAcc* correlates positively with expert variability. It requires the definition of true invalid predictions, TI, as the voxels that are uncertain within in the area of disagreement (uncertain and disagreement) and true non-invalid predicitons, TU, as the voxels that are certain in the area of agreement (certain and agreement). Similarly to conventional accuracy, *DisAcc* can be written as $DisAcc = \frac{TI+TU}{N}$, normalised by the total number of voxels N.

4.2.1 Description of Lung CT dataset

We use the LIDC-IDRI [AIMB+11, WZL+17, HM16] dataset for training and testing. This CT dataset contains images of lung nodules and their delineations from four independent expert observers.

We resample data to an isotropic volume resolution of $1 \times 1 \times 1mm^3$. We use 700 patients as a training dataset and 175 patients as a test set for performance evaluation. We crop each volume at the center of the nodule position and produce volumes of $128 \times 128 \times 128$. For the evaluation of the method we use the Dice score as a metric of volume overlap.

4.3 Evaluation and Results

Correlation of Z_S and Z_{var} and Dice score: We analyze correlation between Z_{var} and Z_S (Sect. 4.2) and the actual Dice score in Figure 4.3. Dice score is computed between the absolute

ground truth (average of 4 annotators) and the predicted segmentation. In Figure 4.3(a,b,e,f) we observe linear negative correlation (statistically significant, $p < 0.001$) between Z_{var}, Z_S and the segmentation performance for both networks. Higher negative correlation is observed for DUNet between Z_{var} and Dice score (Figure 4.3e, $\rho = -0.75$) and between Z_S and Dice score (Figure 4.3f, $\rho = -0.67$). There are some cases (10 cases) in both methods that produce uncertainty scores that are not representative for the segmentation quality. Although these nodules do not have any special visual characteristics, the model produces high Dice scores with high uncertainty scores. In Fig 4.3 c, d and g, h the distributions of uncertainty scores are plotted for two different groups of segmentations. Successful segmentations have been empirically defined as those where the Dice score is ≥ 0.80 and unsuccessful segmentations with Dice scores ≤ 0.65. Thus, a threshold for the uncertainty score, which divides the two groups of segmentations can be defined as the intersection of the two distributions, which is close to 0.25.

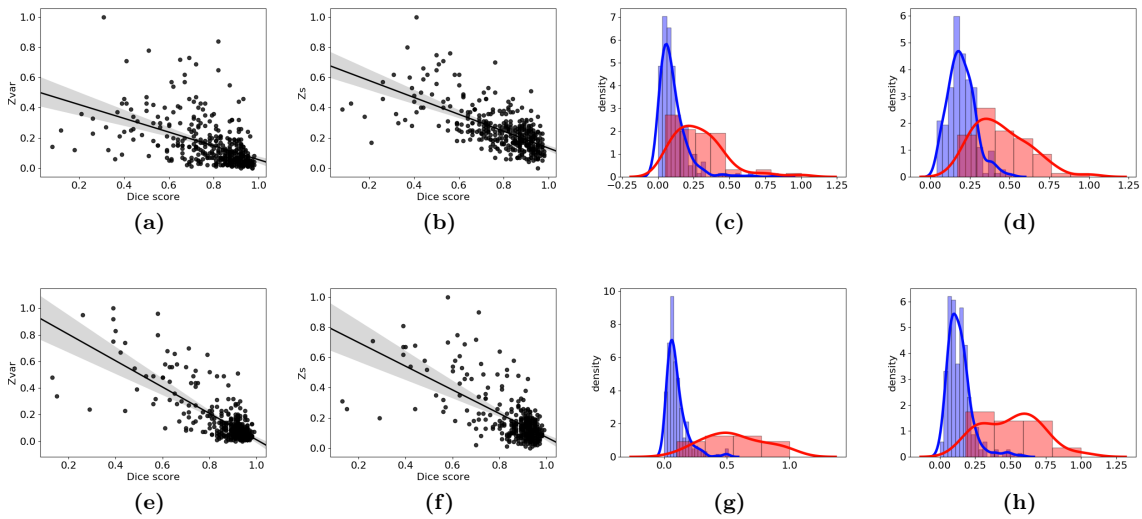
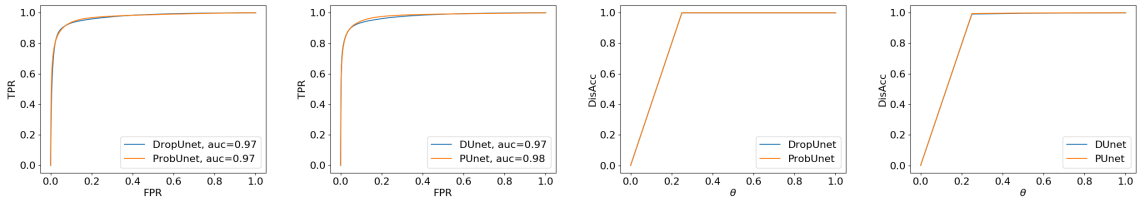


Figure 4.3: Scatter plots of correlation between Dice score and uncertainty scores and probability density function (pdf) plots for both networks. Top row: PUNet. Bottom row: DUNet. Correlation between Dice score and Z_{var} and Z_S respectively: (a-b) PUNet and (e-f) for DUNet. Probability density function (PDF) for values of Z_{var} (and Z_S) of samples whose Dice scores is between 0.80 and 0.95 (blue) and the samples that their Dice scores is lower than 0.65 (red). (c-d) for PUNet and (g-h) for DUNet.

Inter-observer variability vs. segmentation uncertainty: As a naïve baseline we evaluate a convolutional regressor network to predict the annotator variance directly from the volumes. The regressor consists of 5 (convolution-max pooling) layers which are followed by a global

average pooling (GAP) layer to predict $Z_{var} \in [0, 1]$ directly. Mean square error between prediction and ground truth of variance among annotators is minimised during training. The performance of this approach is limited with a mean square error of 0.22 ± 0.0012. To evaluate TPR (Eq. 4.10) and FPR (Eq. 4.11) we compute ROC curves for each network as shown in Figure 4.4 and evaluate *DisAcc* for a range of thresholds. The ROC curves of FPR and TPR (a-b) of both networks are quite similar with the best result for predictive entropy (Eq. 4.6) as uncertainty metric and PUNet with $AUC = 0.98$. Comparing *DisAcc* (c-d) for 5 different thresholds[1] $\theta \in [0, 1]$ for both networks and *DisAcc* reaches 0.99 for $\theta = 0.2$ and then remains stable.



(a) Variance **(b)** Entropy **(c)** Variance **(d)** Entropy

Figure 4.4: ROC curves and *DisAcc* plots using predictive Entropy and σ^2 for both probabilistic networks

Qualitative analysis of inter-observer variability and segmentation uncertainty. Finally, we present a qualitative comparison between segmentation uncertainty and human uncertainty. In Figure 4.5, the uncertainty maps of two lung nodules are presented. The uncertainty maps, which are based on Softmax, predictive entropy, variance and mutual information are compared to an expert-based uncertainty map (annotators' entropy). As can be noticed in the Figure, humans as well as automated estimates for segmentation uncertainty, are greater in the borderline/margin of the nodules.

4.4 Discussion

Using probabilistic 3D segmentation networks, we examine the relationship between segmentation uncertainty and segmentation performance. We explore to which extent human expert inter-observer variability can effect and correlate with voxel-wise segmentation uncertainty. We

[1] It is computed in the range between the minimum and maximum values of the derived segmentation uncertainty.

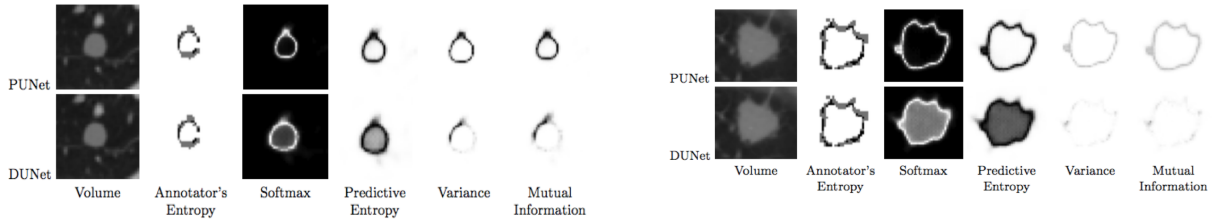


Figure 4.5: Uncertainty maps using maximum Softmax (max(M)), Predictive Entropy (Eq. 4.6), Variance (Eq. 4.4) and Mutual Information (Eq. 4.8) using both networks (darker colour, larger value).

present results that show a relationship between segmentation uncertainty and the area of annotator disagreement. In most cases model segmentation uncertainty indicates also likely human disagreement.

However, there are a few limitations in this framework for which further investigation is needed. Although the utilised models prove to be competitive in the task of uncertainty estimation in segmentation, there are still open issues that should be further examined. These include memory requirements (which is often a limiting factor for processing data with higher resolution or using larger batches), the need for finetuned hyper-parameters (e.g weighting factor in the KL divergence term of the loss in PUNet) [MLFCdC$^+$20] and further examination of the impact of added parameter capacity in the PUNet (i.e posterior net in the training stage). In any case, more efficient ways of modeling uncertainty in ambiguous cases should be sought.

Furthermore, human decision-making is made under uncertainty. The processing of collecting this human uncertainty is usually time consuming, labour-intensive and costly. In the above experiments, we consider annotators' disagreement as a binary problem (i.e normal/abnormal voxel) since this kind of information is available on the dataset. However, it is necessary, additionally to this information, to have access to the level of the annotator's confidence for a delineation of a nodule. Thus, the construction of a dataset which will contain not only binary classification information but also confidence interval for the experts' annotations would be beneficial.

Another limitation of our approach is that for a few cases the evaluated uncertainty scores are not a representative metric for how good or bad is a segmentation and it is likely dependant on the used data set. The application of this framework to additional datasets could be useful

in order to extract more intuition about the reasons of this limitation.

Another point which would be useful to examine is the impact of the different label fusion techniques in the estimation of ground truth and segmentation uncertainty.

Finally, uncertainty as perceived by humans might be fundamentally different from model confidence. Although there is evidence that segmentation uncertainty estimates could also capture/include the human disagreement, it is not clear yet at which extent this happens. Understanding better human (perceptual) uncertainty is vital in order to be able to build a framework which could be able to learn and mimic efficiently the human uncertainty.

4.5 Conclusion

Using probabilistic 3D segmentation networks, we examine the relationship between segmentation uncertainty and segmentation performance. We explore to which extent human expert inter-observer variability can effect and correlate with model segmentation uncertainty.

Our results show that both, a U-Net using MC dropout during inference as well as a 3D probabilistic U-Net architecture can quantitatively correlate the posterior segmentation distribution with true uncertainties. We present results that show evidence that there is a relationship between segmentation uncertainty and the area of annotator disagreement. In most cases model segmentation uncertainty indicates also likely human disagreement.

Chapter 5

Conclusion

5.1 Summary

The main objective of this work was to explore the concepts of automatic anomaly detection and uncertainty under the prism of *variability*. Novel machine learning methods, such as deep learning models, have been applied in this study. *Variability* is discussed in the context of human anatomy as well as in the different stages of the medical imaging processing pipeline.

Understanding variability of human anatomy is vital for the detection and identification of pathologies and subsequently for the treatment of disease. Building large datasets for each disease is a very difficult and time consuming process, thus often infeasible. Additionally, in medical imaging, it is easier to obtain data that conforms with normal appearance. Consequently, effective ways to train successful models should be considered in order to build robust and reliable automatic abnormality detection systems. This usually means to train a model utilising only normal samples and to detect anomalies as deviation from normality. Alternative approaches include attempts to mimic the human ability to learn and understand only from few examples, i.e learning to automatically annotate (unlabeled) data using few-shot learning methods.

The integration of human expert knowledge into medical image processing pipelines entails uncertainty, which is mainly quantified through inter-observer variability. Inter-observer variability is mainly reported in the most ambiguous and difficult cases. Cases which might also

be mis-classified or classified with high uncertainty by an automatic deep learning model. For this reason, it is very important, to explore the relationship between segmentation uncertainty, performance and inter-observer variability. Thus, two key research questions were examined in this work:

The first research question (a) is whether normality can be modelled efficiently during the training phase, in such a way so that it would be possible to detect and localise abnormalities, as deviation from normality, during the test phase. In order to examine this, an unsupervised anomaly detection system based on generative networks with attention mechanisms is applied on normal/healthy subjects. The exemplar application for this method is to detect a subtype of Congenital Heart Disease, in fetal cardiac images during ultrasound screening. An anomaly score which is based on reconstruction and localisation capabilities of the model is proposed. Experimental results validate the advantage of the proposed method over the other state-of-the art algorithms for the specific type of anomaly detection. A thorough analysis of the proposed algorithm both qualitatively as well as quantitatively was presented in *Chapter 3*.

The second research question (b) is whether the uncertainty, which is caused by the integration of human knowledge into the medical imaging processing pipeline, could be captured, modelled and correlated with automatically derived model uncertainty during a training and testing phase. This type of uncertainty appears in the delineation contouring as a inter-observer disagreement over the boundaries of an anomaly. In *Chapter 4* an extensive exploration of the relationship between inter-observer variability, automated estimates of segmentation uncertainty and segmentation performance was performed. Two state-of-the art methods were applied in a 3D imaging segmentation task of lung tumors in CT scans. A metric was established in order to examine the correlation between inter-observer variability and segmentation uncertainty. We examine, both qualitatively and quantitatively, to which extent, automatically predicted confidence and uncertainty metrics, disagreement aware metric (which was proposed) and segmentation performance metrics are correlated. As becomes evident in the experiments, in most cases segmentation uncertainty indicates also human disagreement in annotations. However, in few cases the evaluated uncertainty scores are not a representative metric for how good or bad a segmentation is and it is likely dependant on the used data set.

Overall, there is strong evidence that automatic detection of anomalies, using generative net-

works, can be done with relatively high accuracy and minimal supervision. However, testing these methods in larger clinical cohorts as well as prospective testing are required in order to assess the validity and generalizability of the proposed method in a better way. Furthermore, a very significant parameter is the need for quantification of trust in prediction of abnormalities, which remains a future research challenge. The quantification of model confidence should express domain specific uncertainty which is derived within the various medical imaging processing stages. At the same time, model confidence quantification should be a reliable indicator for the acceptance of a model's prediction. This is necessary to be treated effectively, in order to be able to develop deep learning-based models which will be trustworthy and safe for patient diagnosis, before the integration of these systems into the clinical routine.

5.2 Limitations and Future Work

In this section, the limitations of the proposed methods are examined. Furthermore, potential new research directions and suggestions will be discussed.

Regarding the first research question (a), there are a few issues/limitations that should be tackled efficiently.

A key issue for an automatic anomaly detection system which would be utilised in clinical routine is to never miss an anomaly but at the same time to maintain a low number of false alarms. In this direction, in order to further improve the performance of the proposed abnormality detection method, additional modifications could be suggested.

The proposed AD framework is based on normative learning, following an unsupervised approach. Generative models such as GANs or VAEs based models have proved to show promising anomaly detection performance for this type of learning. However, it would also be interesting to examine the potential benefits of the combination of such generative models with other methods. Proposed modifications could include alternative and more effective loss functions or semi-supervised methods where few abnormal cases will be utilised during training. Further improvements also involve more sophisticated data augmentation methods including self-supervised learning methods via auxiliary pretext tasks and more effective attention methods for more interpretable models.

Another important issue that should be addressed, is the quantification of uncertainty during the detection of abnormalities. It is important, in addition to detecting and localising the anomaly, to provide information about the confidence interval of a model's prediction. This is crucial for building a trustworthy anomaly detection system in safety-critical applications. For instance, in the medical imaging context, high uncertainty in a prediction could be a sign that this case is ambiguous or rare. At the next stage, this case could be examined by an oracle (i.e human expert) for better evaluation (i.e human-in-the-loop approach). In this way, the number of missing critical cases will decrease significantly. Uncertainty quantification in anomaly detection could also be useful as additional information in order to define the cut-off threshold to select the anomalies in a more efficient and meaningful way, in the medical context.

Another very interesting research path which should be investigated together with uncertainty estimation and anomaly detection is interpretability. Concurrently with the effort to move away from the ("black-box") point-estimate predictions (as it was described previously), it is also important to understand better the model's decision process. It could be very beneficial to be able to determine which features contribute most in the model's prediction. The model's outcome would be better understood (by clinicians as well) and along with uncertainty estimation the final decision-making would be much more well-informed. Towards this direction, better and more effective interpretability methods should be proposed. Furthermore, clinical experts should evaluate them for their utility in the clinical routine.

Furthermore, an issue that is under extensive investigation, is the ability of algorithm to distinguish between anomalies and out of distribution samples due to domain shift. Domain shift [GL21] is a common phenomenon in medical imaging because image acquisition is performed at different hospitals/medical sites, using different protocols, from different populations. Domain shift is a common cause for the degradation of a model's performance. As a result, sometimes out-of distribution samples due to domain shift are flagged wrongfully as an anomaly by an automatic abnormality detection system.

The majority of anomaly detection methods in medical imaging are applied on the images and distinguish effectively normal from abnormal anatomy. An alternative approach involves the segmentation of the images at the first stage, and then the application of anomaly detection method on the segmentations. This approach might be useful when applied to specific organs.

For instance, in images with highly variable noisy context, it could be useful to initially segment the images and then detect and localise anomalies. However, if the organ e.g a heart is so unusual or abnormal, the segmentation model might fail. Furthermore, an extra cost and effort is introduced (e.g annotations/labels for ground truth segmentation maps) in order to build a robust (supervised) segmentation model. On the contrary, the utilisation of images in the training process is a less time-consuming process, minimising both expert intervention and image pre-processing. In any case, it would be interesting to examine whether segmentations can be used as complementary information for training an anomaly detection model.

The extension of the proposed method to full video sequence is also a very interesting research direction. This is also a prerequisite for the successful transfer of this framework to real-time application in clinical practice, which is the ultimate goal in this specific research area.

Further extensions include testing our method for other congenital heart disease types as well as the application of the proposed framework on a variety of image modalities such as CT scans or X-Rays. In both cases, the generalisation ability of the proposed method could be validated.

Finally, another extension for this approach concerns the scenario that the abnormalities belong to more than one (different) underlying diseases. Approaches, such as the one proposed in this work, can detect and localise anomalies, however they cannot classify them into a specific disease category. For this purpose, effective deep learning based approaches should be examined and attempted.

There are also research paths that should be extensively investigated in relationship with the examination of the second research challenge (b).

First of all, the development of an "ideal" dataset which will contain annotations from multiple experts and at the same time will provide information about the confidence level of each expert for their annotation would be beneficial. It could be very helpful for a more efficient training of the proposed models and the better understanding of the human perceptual uncertainty.

Despite the success of the existing probabilistic models which produce plausible segmentations (like the ones used in *Chapter 4*), further work needs to be done to improve these models so that they are able to carry the inter-rater variability through to the model's prediction in a more efficient and robust way.

Another very interesting research path for exploration is the integration of the evaluated metrics into clinical quality control (e.g quality of image annotation) or for example into an active learning framework, where 'uncertain' parts of segmentations will be re-processed by a human.

The application of the above methods to other application domains, such as optical coherence tomography, where images suffer from inter-rater variability, could be a also a future research direction.

Finally, an overview of AD methods was presented in *Chapter 2*. This state-of-the-art survey includes both pre-deep learning era methods and deep learning-based methods. An overview like this does not exist in literature yet. The works were discussed based on the conventional pattern recognition or deep learning method that was utilised in each work. However, it would be very interesting to compare the methods also based on the application domain. For instance, a unifying overview of the available AD methods in Ultrasound images solely, would be very interesting in order to explore more efficiently not only the merits and the weaknesses of each method but also the role that acquisition modality plays in medical image analysis. In any case, apart from the theoretical presentation and discussion of the methods, the demonstration of experimental results for medical datasets could be a very interesting future direction.

5.3 Publications

This thesis is mainly based on the following works:

- E. Chotzoglou, T. Day, J. Tan, J. Matthew, D. Lloyd, R. Razavi, J. M. Simpson, B. Kainz. *Learning normal appearance for fetal anomaly screening: Application to the unsupervised detection of Hypoplastic Left Heart Syndrome* In Journal of Machine Learning for Biomedical Imaging (MELBA), 2021.
- E. Chotzoglou, S. Budd, T. Day, J. Simpson, B. Kainz. *Unsupervised detection of Hypoplastic Left Heart Syndrome in fetal screening* Abstract in Workshop on Medical Imaging meets Neurips (MedNeurIPS 2020) at 34th Conference on Neural Information Processing Systems (NeurIPS 2020).
- E. Chotzoglou, B Kainz. *Exploring the Relationship Between Segmentation Uncertainty, Segmentation Performance and Inter-observer Variability with Probabilistic Networks* In LABELS/HAL-MICCAI/CuRIOUS International Workshop, MICCAI, 2019

The following work has not been presented in this thesis, however, it inspired the study of anomaly detection applications in medical imaging.

- R. Holland, U. Patel, P. Lung, E. Chotzoglou, B. Kainz. *Automatic Detection of Bowel Disease with Residual Networks.* In PRIME International Workshop, MICCAI, 2019.

Bibliography

[AAAB19] S. Akcay, A. Atapour-Abarghouei, and T.P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *Asian Conference on Computer Vision (ACCV)*, pages 622–637, 2019.

[AAK18] T. Amarasinghe, A. Aponso, and N. Krishnarajah. Critical analysis of machine learning based approaches for fraud detection in financial transactions. *2018 International Conference on Machine Learning Technologies*, 2018.

[AB18] M. Ayhan and P. Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. *International conference on Medical Imaging with Deep Learning*, 2018.

[ABA06] M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538, 2006.

[ABK+00] S. Albrecht, J. Busch, M. Kloppenburg, F. Metze, and P. Tavan. Generalized radial basis function networks for classification and novelty detection: self-organization of optimal bayesian decision. *Neural Networks*, 13(10):1075–93, 2000.

[Abr18] N. Abroyan. Neural networks for financial market risk classification. *Frontiers in Signal Processing*, 2018.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. arXiv preprint *arXiv:1701.07875*, 2017.

[ACD+98] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study. *Topic Detection and Tracking: Event-Based Information Organization*, page 1–16, 1998.

[ACZ+20] R. Arnaout, L. Curran, Y. Zhao, J. Levine, E. Chinn, and A. Moon-Grady. Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning. *medRxiv*, 2020.

[ADE20] A. Aldweesh, A. Derhab, and A.Z Emam. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, 189:105124, 2020.

[Aey91] D. Aeyels. On the dynamic behavior of the novelty detector and the novelty filter. *Analysis of Controlled Dynamical Systems*, 8:1–10, 1991.

[AF02] M. Augusteijn and B. Folkert. Neural network classification and novelty detection. *International Journal of Remote Sensing*, 23(14):2891–2902, 2002.

[AFR97] E. Aleskerov, B. Freisleben, and B. Rao. CARDWATCH: a neural network based database mining system for credit card fraud detection. *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering, CIFEr 1997, New York City, USA, March 24-25, 1997*, pages 220–226, 1997.

[Aga07] D. Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems*, 11(1):29–44, 2007.

[AGAPN18] J. Ander Gómez, J. Arévalo, R. Paredes, and J. Nin. End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters*, 105:175–181, 2018.

[AGT16] Y. Ando, H. Gomi, and H. Tanaka. Detecting fraudulent behavior using recurrent neural networks. *In Computer Security Symposium*, 2016.

[AHAA+19] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily. Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images. *Sensors*, 19, 2019.

[AHS18] A. Anvari, E. F. Halpern, and A. E. Samir. Essentials of statistical methods for assessing reliability and agreement in quantitative imaging. *Academic radiology*, 25(3):391–396, 2018.

[AIMB+11] S. G Armato III, G. McLennan, L. Bidaut, M. F McNitt-Gray, C. R Meyer, A. P Reeves, B. Zhao, D. R Aberle, C. I Henschke, E. A Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2):915–931, 2011.

[AJBL20] Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening. *Medical Image Analysis*, 60:101618, 2020.

[AKA17] K. Anand, J. Kumar, and K. Anand. Anomaly detection in online social network: A survey. *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 456–459, 2017.

[AKA+20] M.S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724, 2020.

[AMI16] M. Ahmed, A.N Mahmood, and Md. R. Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[AMSCK17] V. Alex, K. P. Mohammed Safwan, S.S. Chennamsetty, and G. Krishnamurthi. Generative adversarial networks for brain lesion detection. *SPIE Medical Imaging*, 10133, 2017.

[ANS19] E. Aliotta, H. Nourzadeh, and J. Siebers. Quantifying the dosimetric impact of organ-at-risk delineation variability in head and neck radiation therapy in the context of patient setup uncertainty. *Physics in medicine and biology*, 64(13), 2019.

[AoR21] Royal Australian and New Zealand College of Radiologists. Quality guidelines for volume delineation in radiation oncology. 2021.

[APCC19] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara. Latent space autoregression for novelty detection. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 481–490, 2019.

[APH+21] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U-R. Acharya, V. Makarenkov, and S. Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[ARA+16] A. Alberts, M. Rempfler, G. Alber, T. Huber, J. Kirschke, C. Zimmer, and B. H. Menze. Uncertainty quantification in brain tumor segmentation using crfs and random perturbation models. *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 428–431, 2016.

[ASF+19] M. Ahmadi, M. Sabokrou, M. Fathy, R. Berangi, and E. Adeli. Generative adversarial irregularity detection in mammography images. *Predictive Intelligence in Medicine*, pages 94–104, 2019.

[AST+21] Hicks S. A., I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa. On evaluation metrics for medical applications of artificial intelligence. *medRxiv*, 2021.

[BB04] S. Basu and R. Bilenko. A probabilistic framework for semi- supervised clustering. *in Proceedings of the 10th ACM Interna- tional Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 59–68, 2004.

[BCKW15] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. *Proceedings of the 32nd International Conference on Machine Learning*, 37:1613–1622, 2015.

[BD20] B. Barz and J. Denzler. Deep learning on small datasets without pre-training using cosine loss. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1360–1369, 2020.

[BDD+01] A. Bronstein, J. Das, M. Duro, R. Friedrich, G. Kleyner, M. Mueller, S. Singhal, and I. Cohen. Self-aware services: using bayesian networks for detecting anomalies in internet-based services. *2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings. Integrated Network Management VII. Integrated Management Strategies for the New Millennium (Cat. No.01EX470)*, pages 623–638, 2001.

[BDGA02] J. M. Barboza, N. K. Dajani, L. G. Glenn, and T. L. Angtuaco. Prenatal diagnosis of congenital cardiac anomalies: A practical approach using two basic views. *RadioGraphics*, 22(5):1125–1138, 2002.

[BDW+20] C. Baur, S. Denner, B. Wiestler, S. Albarqouni, and N. Navab. Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. *arXiv preprint*, arXiv:2004.03271, 2020.

[Bez81] J.C. Bezdek. Pattern recognition with fuzzy objective function algorithms. *Kluwer Academic Publishers, Norwell, MA, USA*, 1981.

[BFD+96] P. Barson, S. Field, N. Davey, , G. McAskey, and R. Frank. The detection of fraud in mobile phone networks. *Neural Network World*, 6(4):477–484, 1996.

[BFN05] N. Benamrane, A. Fréville, and R. Nekkache. A hybrid fuzzy neural networks for the detection of tumors in medical images. *American Journal of Applied Sciences*, 2:892–896, 2005.

[BGS08] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[BGW+20] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pages 718–727, 2020.

[BH99] R. Bolton and D. Hand. Unsupervised profiling methods for fraud detection. *Credut Score and Credit Control VII*, 1999.

[BHMY99] L.D Baker, T. Hofmann, A.K Mccallum, and Y. Yang. A hierarchical probabilistic model for novelty detection in text. *NIPS*, 1999.

[Bis93] C.M Bishop. Novelty detection and neural network validation. *International Conference on Artificial Neural Networks (ICANN)*, pages 789–794, 1993.

[Bis99] C. M. Bishop. Bayesian pca. *in Advances in Neural Information Processing Systems*, page 382–388, 1999.

[Bis06] C. M. Bishop. Pattern recognition and machine learning. *Springer-Verlag*, 2006.

[BIT+17] H.HWJ Bosman, G. Iacca, A. Tejada, H.J Wörtche, and A. Liotta. Spatial anomaly detection in sensor networks using neighborhood information. *Information Fusion*, 33:41–56, 2017.

[BKC17] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (12):2481–2495, 2017.

[BKCTC+19] C. F. Baumgartner, K. C. Kerem C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Urs J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 119–127, 2019.

[BKM+17a] C F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L M. Koch, B. Kainz, and D. Rueckert. Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017.

[BKM17b] D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[BKNS00] MM. Breunig, H-P. Kriegel, RT. Ng, and J. Sander. LOF: identifying Density-Based Local Outliers. *Proceedings of the 2000 ACM SIGMOD International*

Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA, pages 93–104, 2000.

[BL88] D.S. Broomhead and D. Lowe. Radial basis function, multi-variable functional interpolation and adaptive networks. *Royal Signals and Radar Establishment Malvern, Techniqal Report*, 1988.

[BLH99] R.W. Brause, T.S Langsdorf, and H.M Hepp. Neural data mining for credit card fraud detection. *11th IEEE International Conference on Tools with Artificial Intelligence, ICTAI '99, Chicago, Illinois, USA, November 8-10, 1999*, pages 103–106, 1999.

[BMG+10] M. Bennasar, J.M. Martínez, O. Gómez, J. Bartrons, A. Olivella, B. Puerto, and E. Gratacós. Accuracy of four-dimensional spatiotemporal image correlation echocardiography in the prenatal diagnosis of congenital heart defects. *Ultrasound in Obstetrics and Gynecology, 36(4)*, pages 458–464, 2010.

[BMHK+17] L. Banjanovic-Mehmedovic, A. Hajdarevic, M. Kantardzic, F. Mehmedovic, and I. Dzananovic. Neural network-based data-driven modelling of anomaly detection in thermal power plant. *Automatika*, 58(1):69–79, 2017.

[BMVT20] B. Bozorgtabar, D. Mahapatra, G. Vray, and J-P Thiran. Salad: Self-supervised aggregation learning for anomaly detection on x-rays. *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pages 468–478, 2020.

[BRJ+18] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, and et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint*, arXiv:1811.02629, 2018.

[BS73] D. Ballard and J. Sklansky. Tumor detection in radiographs. *Computers and Biomedical Research*, 6(4):299–321, 1973.

[BSG+06] J. Branch, B. Szymanski, C. Giannella, Ran Wolff, and H. Kargupta. In-network outlier detection in wireless sensor networks. *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, 2006.

[BWAN18] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *International MICCAI Brainlesion Workshop*, pages 161–169, 2018.

[BWAN19] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Fusing unsupervised and supervised deep learning for white matter lesion segmentation. *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, 102:63–72, 2019.

[BWAN20] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1905–1909, 2020.

[BWJ01] D. Barbará, N. Wu, and S. Jajodia. Detecting novel network intrusions using bayes estimators. *Proceedings of the First SIAM International Conference on Data Mining, SDM*, pages 1–17, 2001.

[BYW+20] C. Bian, C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, and Y. Zheng. Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64:101732, 2020.

[ÇAL+16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, pages 424–432, 2016.

[CB00] C. Campbell and K.P Bennett. A linear programming approach to novelty detection. *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 395–401, 2000.

[CBD+90] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems 2*, page 396–404, 1990.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

[CC10] W. Chang and J. Chang. Using clustering techniques to analyze fraudulent behavior changes in online auctions. *2010 International Conference on Networking and Information Technology*, pages 34–38, 2010.

[CEH18] A. Chouiekh and E.H.I El Haj. Convnets for fraud detection analysis. *Procedia Computer Science*, 127:133–138, 2018.

[CEWB97] K.C Cox, S.G Eick, G.J Wills, and R.J Brachman. Visual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery*, 1(2):225–231, 1997.

[CF03] A. Chiu and A. Fu. Enhancements on local outlier detection. *Proceedings of the 7th International Database Engineering and Applications Symposium, IEEE*, page 298–307, 2003.

[CFG14] T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. *Proceedings of the 31st International Conference on Machine Learning*, 32(2):1683–1691, 2014.

[CGP+20] E. Chiou, F. Giganti, S. Punwani, I. Kokkinos, and E. Panagiotaki. Harnessing uncertainty in domain adaptation for mri prostate lesion segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 510–520, 2020.

[Cha15] R. Chadwick. Normality as convention and as scientific fact. *Handbook of the Philosophy of Medicine*, pages 1–12, 2015.

[Cha18] L. R. Chai. Uncertainty estimation in bayesian neural networks and links to interpretability. *University of Cambridge*, 2018.

[CHHC20] T. Cao, C. Huang, D. Y-T. Hui, and J.P. Cohen. A benchmark of medical out of distribution detection. *arXiv preprint*, arXiv:2007.04250, 2020.

[CHP+20] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana. Uncertainty estimation in deep neural networks for dermoscopic image classification. *2020*

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 3211–3220, 2020.

[CHRP07] C Chew, JL Halliday, MM Riley, and DJ Penny. Population-based study of antenatal detection of congenital heart disease by ultrasound examination. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 29(6):619–624, 2007.

[CJ18] H. Cho and E. Jang. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint*, arXiv:1810.01392, 2018.

[CK18] X. Chen and E. Konukoglu. Unsupervised Detection of Lesions in Brain MRI using Constrained Adversarial Autoencoders. *MIDL 2018 Medical Imaging with Deep Learning*, 2018.

[CKPB18] J. Cowton, I. Kyriazakis, T. Plötz, and J. Bacardit. A combined deep learning gru-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors. *Sensors*, 18(8):2521, 2018.

[CKPR99] M. Chen, T. Kanade, D. Pomerleau, and H.A Rowley. Anomaly detection through registration. *Pattern Recognition*, 32(1):113–128, 1999.

[CL19] Y. Chen and S. Li. A lightweight anomaly detection method based on svdd for wireless sensor networks. *Wireless Personal Communications*, 105:1235–1256, 2019.

[CLT+18] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie. Deep learning and its applications in biomedicine. *Genomics, Proteomics & Bioinformatics*, 16(1):17–32, 2018.

[CMC17] R. Chalapathy, A.K Menon, and S. Chawla. Robust, deep and inductive anomaly detection. *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017*, pages 36–51, 2017.

[CMHN02] P. Crook, S. Marsland, G. Hayes, and U. Nehmzow. A tale of two filters-online novelty detection. *in Proceedings of the IEEE International Conference on Robotics and Automation*, 4:3894–3899, 2002.

[CMK+20] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, and J. Berent. A simple probabilistic method for deep classification under input-dependent label noise. *arXiv preprint*, arXiv:2003.06778, 2020.

[CMM+14] L. Caravatta, G. Macchia, G. Mattiucci, A. Sainato, N.LV. Cernusco, G. Mantello, M. Di Tommaso, M. Trignani, A. De Paoli, G. Boz, M. L. Friso, V. Fusco, M. Di Nicola, A. G. Morganti, and D. Genovesi. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. *Radiation Oncology*, 9, 2014.

[COL+13] Y. Chung, S. Oh, J. Lee, D. Park, HH. Chang, and S. Kim. Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. *Sensors*, 13(10):12929–42, 2013.

[CP+19] X. Chen, , N. Pawlowski, B. Glocker, and E. Konukoglu. Unsupervised lesion detection with locally gaussian approximation. *Machine Learning in Medical Imaging*, pages 355–363, 2019.

[CPGM06] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglaris. Hierarchical anomaly detection in distributed large-scale sensor networks. *11th IEEE Symposium on Computers and Communications (ISCC'06)*, pages 761–767, 2006.

[CPS17] J. Castellini, V. Poggioni, and G. Sorbi. Fake twitter followers detection by denoising autoencoder. *Proceedings of the International Conference on Web Intelligence*, page 195–202, 2017.

[CRJ+17] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, and et. al. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77–102, 2017.

[CRT+18] N Codella, V. Rotemberg, P. Tschandl, ME. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection 2018: A Challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint*, arXiv:1902.03368, 2018.

[CSHB18] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[CSK+07] W. Chiracharit, Y. Sun, P. Kumhom, K. Chamnongthai, C. F. Babbs, and E. J. Delp. Normal mammogram detection based on local probability difference transforms and support vector machines. *IEICE - Transactions on Information and Systems*, page 258–270, 2007.

[CvMG+14] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *arXiv preprint*, arXiv:1406.1078, 2014.

[CYTK20] X. Chen, S. You, K.C. Tezcan, and E. Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis*, 64:101713, 2020.

[CZSB20] G. Carneiro, L. Zorron Cheng Tao Pu, R. Singh, and A. Burt. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis*, 62:101653, 2020.

[DB16] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in Neural Information Processing Systems 29 (NIPS)*, 29, 2016.

[DDCP88] E.R DeLong, D.M DeLong, and D.L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[Dem68] A.P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

[Dep19] S. Depeweg. Modeling epistemic and aleatoric uncertainty with bayesian neural networks and latent variables. *University of Cambridge*, 2019.

[DH02a] I. Diaz and J. Hollmen. Residual generation and visualization for understanding novel process conditions. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02*, 3:2070–2075, 2002.

[DH02b] C. P. Diehl and J. B. Hampshire. Real-time object classification and novelty detection for collaborative video surveillance. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02*, 3:2620–2625, 2002.

[DHLDVU18] S. Depeweg, J-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *Proceedings of the 35th International Conference on Machine Learning*, 80:1184–1193, 2018.

[DJC98] M.J Desforges, P. Jacob, and J. Cooper. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 12:687–703, 1998.

[DL90] J. S. Denker and Y. LeCun. Transforming neural-net output levels to probability distributions. *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, page 853–859, 1990.

[DLX+20] Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, and Y. Shi. Uncertainty-aware training of neural networks for selective medical image segmentation. *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, 121:156–173, 2020.

[DMTV03] M. De Santo, M. Molinara, F. Tortorella, and M. Vento. Automatic classification of clustered microcalcifications by a multiple expert system. *Pattern Recognition*, 36(7):1467–1477, 2003.

[DPD12] O. V. Demler, M. J. Pencina, and Sr D'Agostino, R. B. Misuse of delong test to compare aucs for nested models. *Statistics in medicine*, 31(23):2577–2587, 2012.

[DSDVL02] A. Datta, M. Shah, and N. Da Vitoria Lobo. Person-on-person violence detection in video data. *Object recognition supported by user interaction for service robots*, pages 433–438, 2002.

[dSPBC19] M.L di Scandalea, C. S. Perone, M. Boudreau, and J. Cohen-Adad. Deep active learning for axon-myelin segmentation on histology data. *arXiv preprint*, arXiv:1907.05143, 2019.

[DVF+98] Q. Diheng, B.M Vetter, Wang. F, R. Narayan, S.F Wu, Y.F. Hou, F. Gong, and C. Sargor. Statistical anomaly detection for link-state routing protocols. *Proceedings Sixth International Conference on Network Protocols (Cat. No.98TB100256)*, pages 62–70, 1998.

[DVR+19] L. Deecke, R. Vandermeulen, L. RuffStephan, S. Mandt, and M. Kloft. Image Anomaly Detection with Generative Adversarial Networks. *Machine Learning and Knowledge Discovery in Databases*, pages 3–17, 2019.

[DWH18] KG Dizaji, X. Wang, and H. Huang. Semi-supervised generative adversarial network for gene expression inference. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[EAP+02] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. *Applications of Data Mining in Computer Security. Advances in Information Security*, 6, 2002.

[EBVJvD+17] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J.A.W.M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[ECKK21] E. Erdil, K. Chaitanya, N. Karani, and E. Konukoglu. Task-agnostic out-of-distribution detection using kernel density estimation. *arXiv preprint*, arXiv:2006.10712, 2021.

[ECPUS19] S. Eggenreich, C. Christian Payer, M. Urschler, and D. Stern. Variational inference and bayesian cnns for uncertainty estimation in multi-factorial bone age prediction. *arXiv preprint*, arXiv:2002.10819, 2019.

[Edg87] FY Edgeworth. On discordant observations. *Philosophical Magazine*, 23(5):364–375, 1887.

[EES06] P. F. Evangelista, M. J. Embrechts, and B. K. Szymanski. Taming the curse of dimensionality in kernels and novelty detection. *In Applied Soft Computing Technologies: The Challenge of Complexity, Springer Verlag*, page 431–444, 2006.

[EKN+17] A. Esteva, B. Kuprel, RA Novoa, J. Ko, SM Swetter, HM Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[ERKL16] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.

[ERVOC19] Z. Eaton-Rosen, T. Varsavsky, S. Ourselin, and M. J. Cardoso. As easy as 1, 2...4? uncertainty in counting tasks for medical imaging. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 356–364, 2019.

[FCTZ16] K. Fu, D. Cheng, Y. Tu, and L. Zhang. Credit card fraud detection using convolutional neural networks. *Neural Information Processing*, pages 483–490, 2016.

[FFG+19] A. Filos, S. Farquhar, A. N. Gomez, T.G.J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint*, arXiv:1912.10481, 2019.

[FG20] A. Farzad and T. Aaron Gulliver. Unsupervised log message anomaly detection. *ICT Express*, 6(3):229 – 237, 2020.

[FGG09] O. Freifeld, H. Greenspan, and J. Goldberger. Multiple sclerosis lesion detection using constrained gmm and curve evolution. *International Journal of Biomedical Imaging*, 2009, 2009.

[FHN+16] S. Faghih-Roohi, S. Hajizadeh, A. Núñez, R. Babuska, and B. De Schutter. Deep convolutional neural networks for detection of rail surface defects. *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2584–2589, 2016.

[FI12] A. Fischer and C. Igel. An introduction to restricted boltzmann machines. *in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, 2012.

[FKM+20] T. Fujioka, K. Kubota, M. Mori, Y. Kikuchi, L. Katsuta, E. Kimura, M.and Yamaga, M. Adachi, G. Oda, T. Nakagawa, Y. Kitazume, and U. Tateishi. Efficient anomaly detection with generative adversarial network for breast ultrasound imaging. *Diagnostics*, 10(7), 2020.

[FLHLT19] A.Y.K. Foong, Y. Li, J. M. Hernández-Lobato, and R. E. Turner. 'In-Between' uncertainty in bayesian neural networks. *Workshop on Uncertainty and Robustness in Deep Learning (ICML)*, 2019.

[Flo08] C. M. Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical biochemist Reviews*, 29, 2008.

[FN19] M.B Fadhel and K. Nyarko. Gan augmented text anomaly detection with sequences of deep statistics. *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5, 2019.

[FP99] T. Fawcett and F. Provost. Activity monitoring: noticing interesting changes in behavior. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 53–62, 1999.

[FPS+16] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288, 2016.

[FR03] C. R. Fox and Y. Rottenstreich. Partition priming in judgment under uncertainty. *Psychological Science*, 14(3):195–200, 2003.

[FS10] M. Filippone and G. Sanguinetti. Information theoretic novelty detection. *Pattern Recognition*, 43(3):805–814, 2010.

[FSP+19] U. Fiore, AD. Santis, F. Perla, P. Zanetti, and F. Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479:448–455, 2019.

[FVPT12] V. Fritsch, G. Varoquaux, J.-B. Poline, and B. Thirion. Non-parametric density modeling and outlier detection in medical imaging datasets. *International Workshop on Machine Learning in Medical Imaging (MLMI), Lecture Notes in Computer Science*, 7588:210–217, 2012.

[FXH+18] K. Faust, Q. Xie, D. Han, K. Goyle, Z. Volynskaya, U. Djuric, and P. Diamanidis. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics*, 19:173, 2018.

[FYKP17] A. Fuentes, S. Yoon, S. Kim, and D. Park. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17, 2017.

[GAA+17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and AC. Courville. Improved training of Wasserstein GANs. arXiv preprint *arXiv:1704.00028*, 2017.

[Gal16] Y. Gal. Uncertainty in deep learning. *PhD Thesis, University of Cambridge*, 2016.

[Gam06] M. Gamon. Graph-based text representation for novelty detection. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA*, pages 17–24, 2006.

[GAS05] R. Gwadera, M. Atallah, and W. Szpankowski. Reliable detection of episodes in event sequences. *Knowledge and Information Systems*, 7:415–437, 2005.

[GEK20] M. Gantenbein, E. Erdil, and E. Konukoglu. Revphiseg: A memory-efficient neural network for uncertainty quantification in medical image segmentation. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis - Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020*, 12443:13–22, 2020.

[Ger03] D Gering. Recognizing deviations from normalcy for brain tumor segmentation. *PhD thesis, MIT*, 2003.

[GEY18] I. Golan and R. El-Yaniv. Deep anomaly detection using geometric transformations. *NIPS*, 2018.

[GG16a] Y. Gal and Z Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint*, arXiv:1506.02158, 2016.

[GG16b] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint*, arXiv:1506.02142, 2016.

[GGG+19] F. C. Ghesu, B. Georgescu, E. Gibson, S. Gündel, M. K. Kalra, R. Singh, S.R. Digumarthy, S. Grbic, and D. Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. *arXiv preprint*, arXiv:1906.07775, 2019.

[GH00] Z. Ghahramani and G. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, 2000.

[GIG17] Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1183–1192, 2017.

[GJ14] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 32:1764–1772, 2014.

[GKK+19] S. Garg, K. Kaur, N. Kumar, G. Kaddoum, A.Y Zomaya, and R Ranjan. A hybrid deep learning-based model for anomaly detection in cloud datacenter networks. *IEEE Transactions on Network and Service Management*, 16(3):924–935, 2019.

[GL21] H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *arXiv preprint*, arXiv:2102.09508, 2021.

[GLL+19] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019.

[GLS+16] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using oriented violent flows. *Image and Vision Computing*, 48-49:37–41, 2016.

[GMEK99] S.E Guttormsson, R. J. Marks, M. A. El-Sharkawi, and I. Kerszenbaum. Elliptical novelty grouping for on-line short-turn detection of excited running rotors. *IEEE Transactions on Energy Conversion*, 14(1):16–22, 1999.

[GPAM+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *In Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[GPSW17] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger. On calibration of modern neural networks. *In Proceeding of the 34th International Conference on Machine Learning*, 70:1321–1330, 2017.

[GR94] S. Ghosh and D.L Reilly. Credit card fraud detection with a neural-network. *27th Annual Hawaii International Conference on System Sciences (HICSS-27), January 4-7, 1994, Maui, Hawaii, USA*, pages 621–630, 1994.

[Gru69] F.E. Grubbs. Procedures for detecting outlying observations in samples. *technometrics*, 11:1–21, 1969.

[GRUG17] A.N. Gomez, M. Ren, R. Urtasun, and R.B. Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in Neural Information Processing Systems 30*, page 2214–2224, 2017.

[GTS20] B. Ghoshal, A. Tucker, and W L Sanghera, B.and Wong. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence*, 2020.

[GWC98] A.K Ghosh, J. Wanken, and F. Charron. Detecting anomalous and unknown intrusions against programs. *Proceedings 14th Annual Computer Security Applications Conference (Cat. No.98EX217)*, pages 259–267, 1998.

[GZZ+20] Y. Gong, Y. Zhang, H. Zhu, J. Lv, H. Cheng, Q. Zhang, Y. He, and S. Wang. Fetal congenital heart disease echocardiogram screening based on dgacnn: Adversarial one-class classification combined with video transfer learning. *IEEE Transactions on Medical Imaging*, 39(4):1206–1222, 2020.

[HA04] V.J Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[HAB+20] K. Hoebel, V. Andrearczyk, A. Beers, J. B. Patel, K. Chang, A. Depeursinge, H. Müller, and J. Kalpathy-Cramer. An exploration of uncertainty information for segmentation quality assessment. *Medical Imaging 2020: Image Processing, (SPIE)*, 11313, 2020.

[Has70] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Haw74] D. M Hawkins. The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346):340–344, 1974.

[Haw80] D. Hawkins. Identification of outliers. *Monographs on Statistics and Applied Probability, Springer Netherlands*, 1980.

[HBH+16] E. Hodo, X. Bellekens, A. Hamilton, P. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson. Threat analysis of iot networks using artificial neural net-

work intrusion detection system. *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6, 2016.

[HBWP13] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[HFLP01] PS Horn, L. Feng, Y. Li, and AJ Pesce. Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47(12):2137–2145, 2001.

[HHGL11] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint*, arXiv:1112.5745, 2011.

[HKF04] V. Hautamaki, I. Karkkainen, and P. Franti. Outlier detection using k-nearest neighbour graph. *17th International Conference on Pattern Recognition (ICPR)*, 3:430–433, 2004.

[HLH19] HLHS. Facts about Hypoplastic Left Heart Syndrome, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/heartdefects/hlhs.html, 2019.

[HLV03] W. Hu, Y. Liao, and V. R. Vemuri. Robust anomaly detection using support vector machines. *In Proceedings of the International Conference on Machine Learning,*, pages 282–289, 2003.

[HLW16] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *arXiv preprint*, arXiv:1608.06993, 2016.

[HM16] M. C Hancock and J. F. Magnan. Lung nodule malignancy classification using only radiologist quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods. *SPIE Journal of Medical Imaging*, 2016.

[HMWJ15] B. J. Holland, J. A. Myers, and C. R. Woods Jr. Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise

prior to planned neonatal cardiac surgery: a meta-analysis. *Ultrasound in Obstetrics and Gynecology*, 45:631 – 638, 2015.

[Hof07] H. Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

[HOT06] G. E. Hinton, S. Osindero, and Y-W Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[HOT+19] Y. Hiasa, Y. Otake, M. Takao, T. Ogawa, N. Sugano, and Y. Sato. Automated muscle segmentation from clinical ct using bayesian u-net for personalized musculoskeletal modeling. *IEEE Transactions on Medical Imaging*, 39:1030–1040, 2019.

[HP02] L. Huang and H. Pashler. Symmetry detection and visual attention: A "binary-map" hypothesis. *Vision research*, 42(11):1421–1430, 2002.

[HRM+20] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. A Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh. Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *arXiv preprint*, arXiv:2007.13559, 2020.

[HS97] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[HSK+12] G.E Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint*, arXiv:1207.0580, 2012.

[HvC93] G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. *In Proceedings of the Sixth Annual Conference on Computational Learning Theory*, page 5–13, 1993.

[HW17] Y. Heryadi and H.K.H.S. Warnars. Learning temporal representation of transaction amount for fraudulent transaction recognition using cnn, stacked lstm, and cnn-lstm. *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2017.

[HWL+18] Z. Han, B. Wei, S. Leung, IB. Nachum, D. Laidley, and S. Li. Automated pathogenesis-based diagnosis of lumbar neural foraminal stenosis via deep multiscale multitask learning. *Neuroinformatics*, 16:325–337, 2018.

[HWS+12] J. Hu, F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. *AMIA, Annual Symposium proceedings*, pages 360–369, 2012.

[HYZ+20] C. Huang, F. Ye, Y. Zhang, Y-F Wang, and Q. Tian. Esad: End-to-end deep semi-supervised anomaly detection. *arXiv preprint*, arXiv:2012.04905, 2020.

[HZ94] G. E. Hinton and R. Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in Neural Information Processing Systems (NIPS)*, pages 3–10, 1994.

[HZRS16] 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*, pages 770–778, 2016.

[Hä90] W. Härdle. Applied nonparametric regression. *Cambridge University Press*, 1990.

[IB09] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[IÇG+18] E. Ilg, Ö. Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox. Uncertainty estimates for optical flow with multi-hypotheses networks. *In: Proceedings of the European Conference on Computer Vision (ECCV)*, page 652–667, 2018.

[Iea19] Jeremy Irvin and et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 2019.

[IGV+18] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos. Detecting and locating gastrointestinal anomalies using deep

learning and iterative cluster unification. *IEEE Transactions on Medical Imaging*, 37(10):2196–2210, 2018.

[IK04] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 440–449, 2004.

[int16] *Accuracy Requirements and Uncertainties in Radiotherapy.* Number 31 in Human Health Series. International Atomic Energy Agency, Vienna, 2016.

[iOG11] Multi-Institutional Target Delineation in Oncology Group. Human-computer interaction in radiotherapy target volume delineation: a prospective, multi-institutional comparison of user input devices. *Journal of Digital Imaging*, 24(5):794–803, 2011.

[ITW18] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *Proceedings of the 35th International Conference on Machine Learning*, 80:2127–2136, 2018.

[IYC+17] J. Inoue, Y. Yamagata, Y. Chen, C.M. Poskitt, and J. Sun. Anomaly detection for a water treatment system using unsupervised machine learning. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1058–1065, 2017.

[IZ18] J. Islam and Y. Zhang. Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics*, 5, 2018.

[IZZE17] P. Isola, J-H. Zhu, T. Zhou, and AA. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[Jag91] A. Jagota. Novelty detection on a very large number of memories stored in a hopfield-style network. *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 2:905, 1991.

[JDGSSC97] J.R J.R. Dorronsoro, F. Ginel, C. Sanchez, and C. Santa Cruz. Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks*, 8(4):827–834, 1997.

[JGZ+18] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, PE. Portier, L. He-Guelton, and O. Caelen. Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100:234–245, 2018.

[JJJ+19] M.H. Jensen, D.R. Jørgensen, R. Jalaboi, M.E. Hansen, and M.A. Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. *In Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019.

[JKR06] D. Janakiram, A. V. U. P. Kumar, and A. M. Reddy V. Outlier detection in wireless sensor networks using bayesian belief networks. *2006 1st International Conference on Communication Systems Software Middleware*, pages 1–6, 2006.

[JME+18] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, and M. Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 682–690, 2018.

[Jol02] I.T. Jolliffe. Principal component analysis. *Springer Series in Statistics*, 2002.

[JR19] A. Jungo and M. Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 11765:48–56, 2019.

[JS02] S. Jakubek and T. Strasser. Fault-diagnosis using neural networks with ellipsoidal basis functions. *Proceedings of the 2002 American Control Conference*, 5:3846–3851, 2002.

[JSR+20] A. Jalalifar, H. Soliman, M. Ruschin, A. Sahgal, and A. Sadeghi-Naini. A brain tumor segmentation framework based on outlier detection using one-class support vector machine. *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1067–1070, 2020.

[JZK03] J. Jiang, C. Zhang, and M. Kamel. Rbf-based real-time hierarchical intrusion detection systems. *Proceedings of the International Joint Conference on Neural Networks*, 2:1512–1516, 2003.

[KA51] S. Kullback and Leibler R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[KBC17] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017.

[KBM+20] E. Klang, Y. Barash, RY. Margalit, S. Soffer, O. Shimon, A. Albshesh, S. Ben-Horin, MM. Amitai, R. Eliakim, and U. Kopylov. Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy. *Gastrointestinal Endoscopy*, 91:606–613, 2020.

[KCL21] KCL. *King's College London*, 2021.

[KCN+20] D. Kimura, S. Chaudhury, M. Narita, A. Munawar, and R. Tachibana. Adversarial Discriminative Attention for Robust Anomaly Detection. *IEEE Winter Conference on Applications of Computer Vision, WACV*, pages 2161–2170, 2020.

[KG17] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30:5574–5584, 2017.

[KHD19] G. Kwon, C. Han, and R.V Daeshik. Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part III*, pages 118–126, 2019.

[KK93] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

[KLK+16] J.F.P. Kooij, M.C. Liem, J.D Krijnders, T.C. Andringa, and D.M. Gavrila. Multi-modal human aggression detection. *Computer Vision and Image Understanding*, 144:106–120, 2016.

[KLLH07] E. Keogh, J. Lin, SH Lee, and H. V Herle. Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems*, 11:1–27, 2007.

[KLN+17] K. Kamnitsas, C. Ledig, V. FJ Newcombe, J. P Simpson, A. D Kane, D. K Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[KMRV03] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur. Bayesian event classification for intrusion detection. *Proceedings of the 19th Annual Computer Security Applications Conference*, page 14, 2003.

[KPNK+19] M. Kull, M. Perelló-Nieto, M. Kängsepp, H. Telmo de Menezes e Silva Filho, Song, and P.A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *In Advances in Neural Information Processing Systems*, pages 12295–12305, 2019.

[KRPM+18] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R Ledsam, K. Maier-Hein, SM A. Eslami, D. J. Rezende, and O. Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 6965–6975, 2018.

[KSFF17] M. Kull, T. M. Silva Filho, and P. Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052–5080, 2017.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *in Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1097–1105, 2012.

[KSU+19] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada. Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):212–224, 2019.

[KSW15] D.P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems 28 (NIPS)*, 28, 2015.

[KT07] M. Karnan and K. Thangavel. Automatic detection of the breast border and nipple position on digital mammograms using genetic algorithm for asymmetry approach to detection of microcalcifications. *Computer Methods and Programs in Biomedicine*, 87(1):12–20, 2007.

[KTE+19] Y. Kawaguchi, R. Tanabe, T. Endo, K. Ichige, and K. Hamada. Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 865–869, 2019.

[KTR+17] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

[Kul59] S. Kullback. Information theory and statistics. *John Wiley & Sons*, 1959.

[KW14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR*, 2014.

[KW+20] Y. Kwon, , J-H. Won, BJ Kim, and M. Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

[Kwa08] N. Kwak. Principal component analysis based on l1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1672–1680, 2008.

[KWAP17] R. Kannan, H. Woo, C.C Aggarwal, and H. Park. Outlier detection for text data : An extended version. *2017 SIAM Data Mining Conference*, pages 489–497, 2017.

[KY18] S. Khan and T. Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241 – 265, 2018.

[KZ17] Z. Kazemi and H. Zarrabi. Using deep networks for fraud detection in the credit card transactions. *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, pages 630–633, 2017.

[LASA+17] C. Leibig, V. Allken, M. Seçkin Ayhan, P. Berens, and S. Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7, 2017.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *in Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LC17] Y. Liu and S. Chawla. Social media anomaly detection: Challenges and solutions. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, page 817–818, 2017.

[LCCC16] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, page 1788–1794, 2016.

[LCD05] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, page 217–228, 2005.

[LCED20] H. Liu, C. Cui, D.J. Englot, and B.M. Dawant. Uncertainty estimation in medical image localization: Towards robust anterior thalamus targeting for deep brain stimulation. *arXiv preprint*, arXiv:2011.02067, 2020.

[LCW+17] MH Le, J. Chen, L. Wang, Z. Wang, W. Liu, KT. Cheng, and X. Yang. Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks. *Physics in Medicine and Biology*, 62(16):6497–6514, 2017.

[LCX+20] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 614–623, 2020.

[LDW+20] G. Luo, S. Dong, W. Wang, K. Wang, S. Cao, C.M. Tam, H. Zhang, J. Howey, P. Ohorodnyk, and S. Li. Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification. *Medical Image Analysis*, 59, 2020.

[LGN21] L. H. Lee, Y. Gao, and J. A. Noble. Principled ultrasound data augmentation for classification of standard planes. *Information Processing in Medical Imaging - 27th International Conference, IPMI*, 12729:729–741, 2021.

[LIF+20] M-H Laves, S. Ihler, J. F. Fast, L.A Kahrs, and T. Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, 121:393–412, 2020.

[LIKO19] M-H Laves, S. Ihler, L.A Kahrs, and T. Ortmaier. Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety. *Current Directions in Biomedical Engineering*, 5:223–226, 2019.

[LIO19] M-H Laves, S. Ihler, and T. Ortmaier. Uncertainty quantification in computer-aided diagnosis: Make your model say "i don't know" for ambiguous cases. *Medical Imaging with Deep Learning (MIDL)*, 2019.

[LJP16] L. Liao, W. Jin, and R. Pavel. Enhanced restricted boltzmann machine with prognosability regularization for prognostics and health assessment. *IEEE Transactions on Industrial Electronics*, 63(11):7076–7083, 2016.

[LKFH05] J. Lin, E.J Keogh, A.W-C Fu, and H.V Herle. Approximations to magic: Finding unusual medical time series. *18th IEEE Symposium on Computer-Based*

Medical Systems (CBMS 2005), 23-24 June 2005, Dublin, Ireland, pages 329–334, 2005.

[LKLHU92] J. A. Leonard, M. Kramer, and L H. Lyle H. Ungar. A neural network architecture that computes its own reliability. *Computers & Chemical Engineering*, 16:819–835, 1992.

[LL19] H. Liu and B. Lang. Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 9:4396, 2019.

[LLC10] Y.-H. Liu, Y.-C. Liu, and Y.-J. Chen. Fast support vector data descriptions for novelty detection. *IEEE Transactions in Neural Networks*, 21(8):1296–1313, 2010.

[LLDL20] X. Li, Y. Lu, C. Desrosiers, and X. Liu. Out-of-distribution detection for skin lesion images with deep isolation forest. *Machine Learning in Medical Imaging - 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings*, pages 91–100, 2020.

[LLG17] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349, 2017.

[LLS+20] D. Lin, K. Lapen, M.V. Sherer, J. Kantor, Z. Zhang, L.M. Boyce, W. Bosch, D. Korenstein, and E.F. Gillespie. A systematic review of contouring guidelines in radiation oncology: Analysis of frequency, methodology, and delivery of consensus recommendations. *International Journal of Radiation, Oncology, Biology, Physics*, 107(4):827–835, 2020.

[LLZ+20] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R.J Radke, and O.I Camps. Towards Visually Explaining Variational Autoencoders. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8639–8648, 2020.

[LMR19] T. LaBonte, C. Martinez, and S. A. Roberts. We know where we don't know: 3d bayesian cnns for credible geometric uncertainty. *arXiv preprint*, arXiv:1910.10793, 2019.

[LN18] T. Luo and S. G. Nagarajan. Distributed anomaly detection using autoencoder neural networks in wsn for iot. *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018.

[LPB17] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6405–6416, 2017.

[LPC+15] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint*, arXiv:1511.06314, 2015.

[LSM+05] C-H Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner. Segmenting brain tumors with conditional random fields and support vector machines. *Computer Vision for Biomedical Image Applications*, pages 469–478, 2005.

[LTA+09] X. A. Li, A. Tai, D. W. Arthur, T. A. Buchholz, S. Macdonald, L. B. Marks, J. M. Moran, L. J. Pierce, R. Rabinovitch, A. Taghian, F. Vicini, W. Woodward, and J. R. White. Variability of target and normal structure delineation for breast-cancer radiotherapy: a RTOG multi-institutional and multi-observer study. *International journal of radiation oncology, biology, physics*, 73(3):944–951, 2009.

[LTMS10] T. Le, D. Tran, W. Ma, and D. Sharma. An optimal sphere and two large margins approach for novelty detection. *in Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–6, 2010.

[LTMS11] T. Le, D. Tran, W. Ma, and D. Sharma. Multiple distribution data description learning algorithm for novelty detection. *Advances in Knowledge Discovery and Data Mining*, 6635:246–257, 2011.

[Lu17] Y. Lu. Deep neural networks and fraud detection. *Project Report, Uppsala Universitet*, 2017.

[LuK00] J. Laurikkala, M. uhola, and E. Kentala. Informal identification of outliers in medical data. *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, 2000.

[LURQ18] S. Latif, M. Usman, R. Rana, and J. Qadir. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. *IEEE Sensors Journal*, 18(22):9393–9400, 2018.

[LX18] Y. Lu and P. Xu. Anomaly detection for skin disease images using variational autoencoder. *arXiv preprint*, arXiv:1807.01349, 2018.

[LYL+18] X. Li, W. Yu, T. Luwang, J. Zheng, X. Qiu, J. Zhao, L. Xia, and Y. Li. Transaction fraud detection using gru-centered sandwich-structured model. *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2018.

[LYN+05] W. Liyang, Y. Yongyi, R.M. Nishikawa, M. N. Wernick, and A. Edwards. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Transactions on Medical Imaging*, 24(10):1278–1285, 2005.

[LZG15] X. Li, F. Zhao, and Y. Guo. Conditional restricted boltzmann machines for multi-label learning with incomplete labels. *in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 38:635–643, 2015.

[LZWJ20] G. Liang, Y. Zhang, X. Wang, and N. Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. *In British Machine Vision Conference (BMVC)*, 2020.

[Mac92] D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 43(3):448–472, 1992.

[Man02] G. Manson. Identifying damage sensitive, environment insensitive features for damage detection. *Proceedings of the IES Conference. Swansea, UK*, 2002.

[MBvH21] S. Mercieca, JSA. Belderbos, and M. van Herk. Challenges in the target volume definition of lung cancer radiotherapy. *Translational lung cancer research*, 10(4):1983–1998, 2021.

[MCF15] Y.-A. Ma, T. Chen, and E. B. Fox. A complete recipe for stochastic gradient mcmc. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2:2917–2925, 2015.

[MDFFF17] S-M Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[MF15] A. Makhzani and B.J Frey. Winner-take-all autoencoders. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2791–2799, 2015.

[MG18] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 7047–7058, 2018.

[MJB+15] BH Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, MA Weber, T. Arbel, BB Avants, N. Ayache, P. Buendia, DL. Collins, N. Cordier, JJ. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, CR. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, KM Iftekharuddin, R. Jena, NM. John, E. Konukoglu, D. Lashkari, JA. Mariz, R. Meier, S. Pereira, D. Precup, SJ Price, TR. Raviv, SM. Reza, M. Ryan, D. Sarikaya, L. Schwartz, HC. Shin, J. Shotton, CA. Silva, N. Sousa, NK. Subbanna, G. Szekely, TJ. Taylor, OM Thomas, NJ Tustison, G. Unal, F Vasseur, M. Wintermark, DH. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transaction on Medical Imaging*, 34(10), 2015.

[MK18] F.S. Mohammadi and A. Kwasinski. Neural network cognitive engine for autonomous and distributed underlay dynamic spectrum access. *arXiv preprint*, arXiv:1806.11038, 2018.

[MKKY18] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral Normalization for Generative Adversarial Networks. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[MKM18] B. A. Mudassar, J. H. Ko, and S. Mukhopadhyay. An unsupervised anomalous event detection framework with class aware source separation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2671–2675, 2018.

[MLB+05] BH. Menze, MP. Lichy, P. Bachert, BM. Kelm, HP. Schlemmer, and FA. Hamprecht. Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors. *NMR in biomedicine*, 15(5):599–609, 2005.

[MLFCdC+20] M. Monteiro, L. Le Folgoc, D. Coelho de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, and B. van der Wilk, M.and Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems (NeurIPS)*, 33(12756-12767), 2020.

[MLKM+18] M. Mario Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs Created Equal? A Large-Scale Study. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pages 698–707, 2018.

[MLY17] S. Min, B. Lee, and S. Yoon. Deep learning in bioinformatics. *Brief Informatics*, 18(5):851–869, 2017.

[MMCS11] J. Masci, U. Meier, D. C. Ciresan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial*

Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I, pages 52–59, 2011.

[MNM+19] A. Mobiny, H. V. Nguyen, S. Moulik, N. Garg, and C.C Wu. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *arXiv preprint arXiv:1906.04569*, 2019.

[MPKM16] S. Mohammadi, A. Perina, H. Kiani, and V. Murino. Angry crowds: Detecting violent events in videos. *Computer Vision–ECCV 2016*, pages 3–18, 2016.

[MR05] A.S Minhas and M.R Reddy. Neural network based approach for anomaly detection in the lungs region by electrical impedance tomography. *Physiological Measurement*, 26(4), 2005.

[MRR+53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[MS16] J R Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint*, arXiv:1612.00390, 2016.

[MSJ+16] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv:1511.05644*, 2016.

[MSS12] A. Mahapatra, N. Srivastava, and J. Srivastava. Contextual anomaly detection in text data. *Algorithms*, 5(4):469–489, 2012.

[MVDM17] A. Munawar, P. Vinayavekhin, and G. De Magistris. Spatio-temporal anomaly detection for industrial robots through prediction in unsupervised feature space. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1017–1025, 2017.

[MWIT+20] A. Mehrtash, W.M. Wells III, C.M Tempany, P. Abolmaesumi, and T. Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39:3868–3878, 2020.

[MXW+20] Y. Mao, F-F Xue, R. Wang, J. Zhang, W-S Zheng, and H. Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pages 529–538, 2020.

[MY00] L.M. Manevitz and M. Yousef. Learning from positive data for document classification using neural networks. *Proceedings of Second Bar-Ilan Workshop on Knowledge Discovery and Learning*, 2000.

[MY02] L.M. Manevitz and M. Yousef. One-class svms for document classification. *The Journal of Machine Learning Research*, pages 139–154, 2002.

[NC03] C.C Noble and D.J. Cook. Graph-based anomaly detection. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2003.

[Nea96] R. M. Neal. Bayesian learning for neural networks. *Lecture Notes in Statistics, Springer Verlag*, 1996.

[NEN+19] A. Norouzi, A. Emami, K. Najarian, N. Karimi, S. Samavi, and S. Soroushmehr. Exploiting uncertainty of deep neural networks for improving segmentation accuracy in mri images. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2322–2326, 2019.

[NGB+20] M. Ng, F. Guo, L. Biswas, S.E Petersen, S.K Piechnik, S. Neubauer, and G. Wright. Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. *arXiv preprint*, arXiv:2012.15772, 2020.

[NHS15] NHS. *Fetal anomaly screening programme: programme handbook June 2015*. Public Health England, 2015.

[Nje08] C. F. Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics*, 33(4):136–140, 2008.

[NJW14] N. Nabizadeh, N. John, and C. Wright. Histogram-based gravitational optimization algorithm on single mr modality for automatic brain lesion detection and segmentation. *Expert Systems with Applications*, 41(17):7820–7836, 2014.

[NKK20] P. Natekar, A. Kori, and G. Krishnamurthi. Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Frontiers in Computational Neuroscience*, 14, 2020.

[NL20] L. Naud and A. Lavin. Manifolds for unsupervised visual anomaly detection. *arXiv preprint*, arXiv:2006.11364, 2020.

[NPAA18] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 11070:655–663, 2018.

[NPF11] S. Ntalampiras, I. Potamitis, and N. Fakotakis. Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Transactions on Multimedia*, 13(4):713–719, 2011.

[NPTTH20] H. Nguyen, K. Phuc Tran, S. Thomassey, and M. Hamad. Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management. *International Journal of Information Management*, 2020.

[NS16] A. Nanduri and L. Sherry. Anomaly detection in aircraft data using recurrent neural networks (rnn). *2016 Integrated Communications Navigation and Surveillance (ICNS)*, 2016.

[NWC+19] CP. Ngo, AA. Winarto, Khor Li Kou C., S. Park, F. Akram, and HK. Lee. Fence GAN: Towards Better Anomaly Detection. arXiv preprint *arXiv:1904.01209*, 2019.

[Obu03] N. A. Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.

[OGIR14] C. O'Reilly, A. Gluhak, M. Imran, and S. Rajasegarar. Anomaly detection in wireless sensor networks in a non-stationary environment. *IEEE Communications Surveys & Tutorials*, 16:1413–1432, 2014.

[OP19] P. Oza and V. M. Patel. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26:277–281, 2019.

[oR17] The Royal College of Radiologists. Radiotherapy target volume definition and peer review-PCR guidance. *London: The Royal College of Radiologists Ref No. BFCO*, 17(2), 2017.

[OSB+19] J. I. Orlando, P. Seeböck, H. Bogunović, S. Klimscha, C. Grechenig, S. Waldstein, B. S. Gerendas, and U. Schmidt-Erfurth. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1441–1445, 2019.

[OWB17] O. Ozdemir, B. Woodward, and A. A Berlin. Propagating uncertainty in multistage bayesian convolutional neural networks with application to pulmonary nodule detection. *arXiv preprint*, arXiv:1712.00497, 2017.

[OYU+19] K. Ouardini, H. Yang, B. Unnikrishnan, M. Romain, C. Garcin, H. Zenati, J.P Campbell, M.F Chiang, J. Kalpathy-Cramer, V. Chandrasekhar, P. Krishnaswamy, and C-S Foo. Towards practical unsupervised anomaly detection on retinal images. *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data - First MICCAI Workshop, DART*, 11795:225–234, 2019.

[PAD18] S. Pidhorskyi, R. Almohsen, and G. Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, pages 6822–6833, 2018.

[PAL04] C. Phua, D. Alahakoon, and V.C.S Lee. Minority report in fraud detection: classification of skewed data. *SIGKDD Explorer Newsletter*, 6(1):50–59, 2004.

[Par62] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[PATB96] T. Parr, S. Astley, C. Taylor, and C. Boggis. Model based classification of linear structures in digital mammograms (automatic detection and model based clas-

sification of anatomically different linear structures in digital mammograms). *Digital Mammography*, 1996.

[PBC+19] A. Pol, V. Berger, G. Cerminara, C. Germain, and M. Pierini. Anomaly detection with conditional variational autoencoders. *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1651–1657, 2019.

[PBHG04] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–283, 2004.

[PCCT14] M. A. F Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[PH18] S.H. Park and K. Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, 2018.

[PKC+19] O. Potvin, A. Khademi, I. Chouinard, F. Farokhian, L. Dieumegarde, I. Leppert, R. Hoge, M. N. Rajah, P. Bellec, and S. Duchesne. Measurement variability following mri system upgrade. *Frontiers in neurology*, 10(726), 2019.

[PKM+12] NM. Pinto, HT. Keenan, LL. Minich, MD. Puchalski, M. Heywood, and LD. Botto. Barriers to prenatal detection of congenital heart disease: a population-based study. *Ultrasound in Obstetrics and Gynecology, 40(4),*, pages 418–425, 2012.

[Pla99] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances In Large Margin Classifiers*, 10(3):61–74, 1999.

[PMB+13] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

[PMD+95] T. Petsche, A. Marcantonio, C. Darken, SJ. Hanson, G.M. Kuhn, and N.I. Santoso. A neural network autoassociator for induction motor failure prediction. *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*, pages 924–930, 1995.

[PML+05] D. Pokrajac, V. Megalooikonomou, A. Lazarevic, D. Kontos, and Z. Obradovic. Applying spatial distribution analysis techniques to classification of 3d medical images. *Artificial Intelligence in Medicine*, 33(3):261–280, 2005.

[PN97] P.A. Porras and P.G. Neumann. Emerald: Event monitoring enabling responses to anomalous live disturbances. *Proceedings of the 20th National Information Systems Security Conference*, 1997.

[PNX19] P. Perera, R. Nallapati, and B. Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2898–2906, 2019.

[POP21] P. Perera, P. Oza, and V. M. Patel. One-class classification: A survey. *arXiv preprint*, arXiv:2101.03064, 2021.

[PP18] P. Perera and V.M. Patel. Learning deep features for one-class classification. *arXiv preprint*, arXiv:1801.05365, 2018.

[PS15] G. Pachauri and S. Sharma. Anomaly detection in medical wireless sensor networks using machine learning algorithms. *Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems*, 70:325 – 333, 2015.

[PY18] A. Pumsirirat and L. Yan. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of Advanced Computer Science and Applications*, 9, 2018.

[PYS+20] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. *2020 IEEE/CFV Conference on Computer Vision and Pattern recognition (CVPR)*, pages 12170–12179, 2020.

[QCSSL09] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D Lawrence. Dataset shift in machine learning. *The MIT Press*, 2009.

[QLC+16] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel. Multiple-instance learning for anomaly detection in digital mammography. *IEEE Transactions on Medical Imaging*, 35(7):1604–1614, 2016.

[QW07] J. Quinn and C. Williams. Known unknowns: novelty detection in condition monitoring. *Pattern Recognition and Image Analysis*, 4477:1–6, 2007.

[RCN+19] A G Roy, S. Conjeti, N. Navab, C. Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

[RCNW18] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger. Inherent brain segmentation quality control from fully convnet monte carlo sampling. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 664–672, 2018.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 9351:234–241, 2015.

[RH18] M. Renström and T. Holmsten. Fraud detection on unlabeled data with unsupervised machine learning. *Thesis, KTH*, 2018.

[RHC+20] J. C. Reinhold, S. He, Y. andHan, Y. Chen, D. Gao, J. Lee, J. L. Prince, and A. Carass. Validating uncertainty in medical image translation. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 95–98, 2020.

[RHH+20] J. C Reinhold, Y. He, S. Han, Y. Chen, D. Gao, J. Lee, J. L. Prince, and A. Carass. Finding novelty with uncertainty. *SPIE Medical Imaging,*, 2020.

[RIB+17] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, R.L. Laird, D.and Ball, and et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint*, arXiv:1712.06957, 2017.

[RKBN19] M. Ravanbakhsh, T. Klein, K. Batmanghelich, and M. Nabi. Uncertainty-driven semantic segmentation through human-machine collaborative learning. *arXiv preprint*, arXiv:1909.00626, 2019.

[RKV+21] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings IEEE*, 109(5):756–795, 2021.

[RLWFM17] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. Variational Approaches for Auto-Encoding Generative Adversarial Networks. arXiv preprint *arXiv:1706.04987*, 2017.

[RMC16] A. Radford, L. Metz, and S Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016,San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[RMW14] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning, PMLR*, 32(2):1278–1286, 2014.

[RMZS19] L Raczkowski, M. Mozejko, J. Zambonelli, and E. Szczurek. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. *Scientific Reports*, 9:14347, 2019.

[RN19] E. Rushe and B.M Namee. Anomaly detection in raw audio using deep autoregressive networks. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3597–3601, 2019.

[RNS18] S. Ranganathan, K. Nakai, and C. Schonbach. Encyclopedia of bioinformatics and computational biology: Abc of bioinformatics. *Elsevier Science Publishers B. V.*, 2018.

[Rob02] S.J. Roberts. Extreme value statistics for novelty detection in biomedical signal processing. *2000 First International Conference Advances in Medical Signal and Information Processing (IEE Conf. Publ. No. 476)*, pages 166–172, 2002.

[Rot04] V. Roth. Outlier detection with one-class kernel fisher discriminants. *in Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, 2004.

[RPA17] P. Ranganathan, C. S. Pramesh, and R. Aggarwal. Common pitfalls in statistical analysis: Measures of agreement. *Perspectives in clinical research*, 8(4):187–191, 2017.

[RPWS08] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical journal*, 50(3):419–430, 2008.

[RS97] R. Ruotolo and C. Surace. A statistical approach to damage detection through vibration monitoring. *Proceedings of the 5th Pan American Congress of Applied Mechanics. Puerto Rico*, 1997.

[RS20] D. Rueckert and J. A Schnabel. Model-based and data-driven strategies in medical image computing. *Proceedings IEEE*, 108(1):110–124, 2020.

[RSDM19] T. Rott Shaham, T. Dekel, and T. Michaeli. SinGAN: Learning a generative model from a single natural image. *In IEEE Conference on Computer Vision (ICCV)*, 2019.

[RSNL+19] E. Rubinstein, M. Salhov, M. Nidam-Leshem, V. White, S. Golan, J. Baniel, H. Bernstine, D. Groshar, and A. Averbuch. Unsupervised tumor detection in dynamic PET/CT imaging of the prostate. *Medical Image Analysis*, 55:27–40, 2019.

[RSNS19] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1896–1904, 2019.

[RST+11] D. Roberge, T. Skamene, R.E. Turcotte, T. Powell, N. Saran, and C. Freeman. Inter- and intra-observer variation in soft-tissue sarcoma target definition. *Cancer Radiotherapy*, 15:421–425, 2011.

[RT94] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

[RVG+18] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep One-Class Classification. *Proceedings of Machine Learning Research*, pages 4393–4402, 2018.

[SAR+19] S. Sedai, B. Antony, R. Rai, K. Jones, H. Ishikawa, J. Schuman, W. Gadi, and R. Garnavi. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, page 282–290, 2019.

[SBW21] S. Sheynin, S. Benaim, and L. Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. *In IEEE Conference on Computer Vision (ICCV)*, 2021.

[SC05] E. J. Spinosa and A.C. Carvalho. Support vector machines for novel class detection in bioinformatics. *Genetics and molecular research : GMR*, 4(3), 2005.

[SCD+17] R R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In ICCV, 2017.*, 2017.

[SCS18] W Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[SCSC03] M-L. Shyu, S-h. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier. *in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03)*, pages 173–179, 2003.

[SCWZ20] H. Shen, J. Chen, R. Wang, and J. Zhang. Counterfeit Anomaly Using Generative Adversarial Network for Anomaly Detection. *IEEE Access*, pages 133051–133062, 2020.

[Sd16] C. Steyn and A. de Waal. Semi-supervised machine learning for textual anomaly detection. *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–5, 2016.

[SDVG19] C. Sakaridis, D. Dai, and L. Van Gool. Semantic nighttime image segmentation with synthetic stylized data, gradual adaptation and uncertainty-aware evaluation. *arXiv preprint arXiv:1901.05946*, 2019.

[SESG+18] U. Schmidt-Erfurth, A. Sadeghipour, BS. Gerendas, SM. Waldstein, and H. Bogunović. Artificial intelligence in retina. *Progress in Retinal and Eye Research*, 67:1–29, 2018.

[SG18] L. Smith and Y. Gal. Understanding measures of uncertainty for adversarial example detection. *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569, 2018.

[SGRP+20] J. Senapati, A. Guha Roy, S. Pölsterl, D. Gutmann, S. Gatidis, C. Schlett, A. Peters, F. Bamberg, and C. Wachinger. Bayesian neural networks for uncertainty estimation of imaging biomarkers. *MICCAI-MLMI 2020 Workshop*, 2020.

[SH06] R. Salakhutdinov and G. Hinton. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[Sha15] G. Shan. Improved confidence intervals for the youden index. *PloS one*, 10(7), 2015.

[SHK18] T. Sweers, T. Heskes, and J. Krijthe. Autoencoding credit card fraud. *Radboud University*, 2018.

[SHL17] L. Shu, X. Hu, and B Liu. Doc: Deep open classification of text documents. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2911–2916, 2017.

[SHMU19] K. Sato, K. Hama, T. Matsubara, and K. Uehara. Predictable uncertainty-aware unsupervised deep anomaly segmentation. *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2019.

[SHN+18] D. Sato, S. Hanaoka, Y. Nomura, T. Takenaga, S. Miki, T. Yoshikawa, N. Hayashi, and O. Abe. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. *Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis*, 2018.

[SK19] A. Sagheer and M. Kotb. Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems. *Scientific Reports*, 9(19038), 2019.

[SKFA18] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially Learned One-Class Classifier for Novelty Detection. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3379–3388, 2018.

[SKS+19] M. Shehata, F. Khalifa, A. Soliman, M. Ghazal, F. Taher, MA. El-Ghar, AC. Dwyer, G. Gimel'farb, RS. Keynton, and A. El-Baz. Computer-aided diagnostic system for early detection of acute renal transplant rejection using diffusion-weighted MRI. *IEEE Transactions on Biomedical Engineering*, 66:539–552, 2019.

[SL05] HE. Solberg and A. Lahti. Detection of outliers in reference distributions: performance of Horn's algorithm. *Clinical Chemistry*, 51:2326–2332, 2005.

[SLD17] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[SLJ+15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[SLY15] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, pages 3483–3491, 2015.

[SM04] S. Singh and M. Markou. An approach to novelty detection applied to the classification of image regions. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):396–407, 2004.

[SMAN19] R. D Soberanis-Mukul, S. Albarqouni, and N. Navab. An uncertainty-driven gcn refinement strategy for organ segmentation. *arXiv preprint*, arXiv:1906.02191, 2019.

[SMH07] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative filtering. *Proceedings of the 24th International Conference on Machine Learning*, page 791–798, 2007.

[Smi21] H. F. Smith. Anatomical variation and clinical diagnosis. *Diagnostics (Basel, Switzerland)*, 11(2):247, 2021.

[Smy94] P. Smyth. Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12:1600–1612, 1994.

[SOS02] A.A Sebyala, T. Olukemi, and L. Sacks. Active platform security through intrusion detection using naive bayesian network for anomaly detection. *Proceedings of the 2002 London Communications Symposium*, 2002.

[SOS+19] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, pages 197 – 207, 2019.

[SOS+20] P. Seeböck, J. I. Orlando, T. Schlegl, S. Waldstein, H. Bogunovic, S. Klimscha, G. Langs, and U. Schmidt-Erfurth. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE Transactions on Medical Imaging*, 39:87–98, 2020.

[SP16] B. Segedin and P. Petric. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? *Radiology and Oncology*, 50(3):254–262, 2016.

[SPS01] C. Spence, L. Parra, and P. Sajda. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01)*, 2001.

[SPST+01] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[SQCF05] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005.

[Sri06] A.N Srivastava. Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques. *2006 IEEE Aerospace Conference*, 2006.

[SRV+20] J. Swiatkowski, K. Roth, B. S. Veeling, L. Tran, J.V. Dillon, J. Snoek, S. Mandt, T. Salimans, R. Jenatton, and S. Nowozin. The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in bayesian neural networks. *arXiv preprint*, arXiv:2002.02655, 2020.

[SSB+17] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer. Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint*, arXiv:1709.05254, 2017.

[SSL+17] D. Sidibé, S. Sankar, G. Lemaître, M. Rastgoo, J. Massich, C.Y Cheung, G.S.W Tan, D. Milea, E. Lamoureux, T.Y Wong, and F. Mériaudeau. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. *Computer Methods and Programs in Biomedicine*, 139:109–117, 2017.

[SSM98] B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[SSW+17] T. Schlegl, P. Seeböck, SM. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised Anomaly Detection with Generative Adversarial Network to Guide Marker Discovery. *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157, 2017.

[SSW+19] T. Schlegl, P. Seeböck, SM. Waldstein, G Langs, and U.. Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, pages 30–44, 2019.

[SWF01] H. Sohn, K. Worden, and C.H Farrar. Novelty detection under changing environmental conditions. *Proceedings of Eighth Annual SPIE International Symposium on Smart Structures and Materials*, 2001.

[SWJR07] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Transactions on Data and Knowledge Engineering*, 19(5):631–645, 2007.

[SWK+19] P. Seeböck, S.M Waldstein, S. Klimscha, H. Bogunovic, T. Schlegl, B.S Gerendas, R. Donner, U. Schmidt-Erfurth, and G. Langs. Unsupervised identification of disease marker candidates in retinal OCT imaging data. *IEEE Transactions on Medical Imaging*, 38(4):1037–1047, 2019.

[SWS+00] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, 12(3):582–588, 2000.

[SWYT03] E. Suzuki, T. Watanabe, H. Yokoi, and K. Takabayashi. Detecting interesting exceptions from medical test data with visual summarization. *Third IEEE International Conference on Data Mining*, pages 315–322, 2003.

[SX14] X. Sun and W. Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

[SZ02] K. Sequeira and M.J. Zaki. Admit: anomaly-based data mining for intrusions. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 386–395, 2002.

[SZ05] A.N Srivastava and B. Zane-Ulman. Discovering recurring anomalies in text reports regarding complex space systems. *2005 IEEE Aerospace Conference*, pages 3853–3862, 2005.

[SZ15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR*, 2015.

[SZE+21] T. B. Smith, S. Zhang, A. Erkanli, D.P. Frush, and H. Samei. Variability in image quality and radiations dose within and across 97 medical facilities. *Journal of Medical Imaging*, 8(5):052105, 2021.

[SZT+19] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. Gotway, Y. Bengio, and J. Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 191–200, 2019.

[TB99] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[TBFD20] N. Tuluptceva, B. Bakker, H. Fedulova, I. Schulz, and D. V. Dylov. Anomaly detection with deep perceptual autoencoders. *arXiv preprint*, arXiv:2006.13265, 2020.

[TCFC02] J. Tang, Z. Chen, A. Fu, and D. Cheung. Enhancing effectiveness of outlier detections for low density patterns. *Advances in Knowledge Discovery and Data Mining*, 2336:535–548, 2002.

[TCSHPFR09] A. Taboada-Crispi, H. Sahli, D. Hernandez-Pacheco, and A. Falcon-Ruiz. Anomaly detection in medical image analysis. *Handbook of Research on Ad-*

vanced Techniques in Diagnostic Imaging and Biomedical Applications, pages 426–446, 2009.

[TD04] D.M.J Tax and R.P.W Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.

[TES+21] M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. F. Azampour, and N. Navab. Rethinking ultrasound augmentation: A physics-inspired approach. *Medical Image Computing and Computer Assisted Intervention–(MICCAI)*, pages 690–700, 2021.

[THC+19] M. Tesařová, E. Heude, G. Comai, T. Zikmund, M. Kaucká, I. Adameyko, S. Tajbakhsh, and J. Kaiser. An interactive and intuitive visualisation method for x-ray computed tomography data of biological samples in 3d portable document format. *Scientific Reports*, 9, 2019.

[THCB95] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. *1995 Fourth International Conference on Artificial Neural Networks*, pages 442–447, 1995.

[THHT98] M. Taniguchi, M. Haft, J. Hollmén, and V. Tresp. Fraud detection in communication networks using neural and probabilistic methods. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998*, 1241–1244, 1998.

[Tip01] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(3):211–244, 2001.

[TKG21] P. Thiagarajan, P. Khairnar, and S. Ghosh. Explanation and use of uncertainty obtained by bayesian neural network classifiers for breast histopathology images. *IEEE Transactions on Medical Imaging*, 2021.

[TL11] L. Torgo and E. Lopes. Utility-based fraud detection. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI*, pages 1517–1522, 2011.

[TL19] M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning, ICML*, 97:6105–6114, 2019.

[TLP+19] A. Tousignant, P. Lemaître, D. Precup, D.L Arnold, and T. Arbel. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, 102:483–492, 2019.

[TLS89] N. Tishby, E. Levin, and S.A. Solla. Consistent inference of probabilities in layered networks:predictions and generalizations. *International Joint Conference on Neural Networks*, 2:403–409, 1989.

[TLZ+18] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich. Precision and recall for time series. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 1924–1934, 2018.

[TMP+20] Y. Tian, G. Maicas, L.Z.C.T. Pu, R. Singh, J. W. Verjans, and G. Carneiro. Few-shot anomaly detection for polyp frames from colonoscopy. *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pages 274–284, 2020.

[TPMO14] J.T Turner, A. Page, T. Mohsenin, and T. Oates. Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection. *2014 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 24-26, 2014*, 2014.

[TRK18] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 2018.

[TSS+19] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *arXiv preprint*, arXiv:1902.03680, 2019.

[TTHX19] Y-X Tang, Y-B Tang, M. Han, and R. M. Xiao, J. Summers. Deep adversarial one-class learning for normal and abnormal chest radiograph classification. *SPIE Medical Imaging, 2019,*, 10950, 2019.

[TWG+17] R. Tanno, D.E Worrall, A. Ghosh, E. Kaden, S.N. Sotiropoulos, A. Criminisi, and D. C. Alexander. Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017.

[TWK+21] R. Tanno, D.E Worrall, E. Kaden, A. Ghosh, F. Grussu, A. Bizzi, S.N Sotiropoulos, A. Criminisi, and D.C Alexander. Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI. *NeuroImage*, 225:117366, 2021.

[TYBHX19] Y-H. Tang, Tang Y-B., M. Han, and R.M Xiao, J.and Summers. Abnormal chest x-ray identification with generative adversarial one-class classifier. *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*, pages 1358–1361, 2019.

[USHE19] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. *International Journal of Computer Assisted Radiology and Surgery*, 14:451–461, 2019.

[UVA20] U. Uddeshya, S. Viswanath, and S.P Awate. QUEST for MEDISYN: Quasi-norm based uncertainty ESTimation for MEDical image SYNthesis. *International Conference on Machine Learning Workshop on Uncertainty and Robustness in Deep Learning (ICML-UDL)*, 2020.

[UVL18] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. It takes (only) two: Adversarial generator-encoder networks. *AAAI*, 2018.

[Vap95] V. N. Vapnik. The nature of statistical learning theory. *Springer-Verlag New York, Inc., New York, NY, USA*, 1995.

[VGT+07] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. *2007*

IEEE Conference on Advanced Video and Signal Based Surveillance, pages 21–26, 2007.

[VJMH16] S.K. Vinod, M.G. Jameson, M. Min, and L.C. Holloway. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiotherapy and Oncology*, 121:169–179, 2016.

[VLMV+01] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, 2001.

[VMJH16] S.K. Vinod, M. Min, M.G. Jameson, and L.C. Holloway. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *Journal of Medical Imaging and Radiation Oncology*, 60(3):393–406, 2016.

[VMT98] H Verrelst, J. Moreau, Y.and Vandewalle, and D. Timmerman. Use of a multilayer perceptron to predict malignancy in ovarian tumors. *Advances in Neural Information Processing Systems*, 10:978–984, 1998.

[VPHC+06] T. Van Phuong, L. X. Hung, S. J. Cho, Y-K Lee, and S. Lee. An anomaly detection algorithm for detecting attacks in wireless sensor networks. *Intelligence and Security Informatics*, pages 735–736, 2006.

[VPNN20] L. Venturini, A.T Papageorghiou, A. J. Noble, and A.I.L. Namburete. Uncertainty estimates as data selection criteria to boost omni-supervised learning. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 12261:689–698, 2020.

[VPSM19] S. Venkataramanan, K-C. Peng, R.V. Singh, and A. Mahalanobis. Adversarial Discriminative Attention for Robust Anomaly Detection. arXiv preprint *arXiv:1911.08616*, 2019.

[vVCR+16] CL. van Velzen, SA. Clur, MEB. Rijlaarsdam, CJ. Bax, E. Pajkrt, MW. Heymans, MN. Bekker, J. Hruda, CJM. de Groot, NA. Blom, and MC. Haak. Prenatal detection of congenital heart disease–results of a national screening

programme. *BJOG: An international journal in Obstetrics and Gynaecology , 123(3)*, pages 400–407, 2016.

[WBJC08] E.A. White, K.K. Brock, D.A. Jaffray, and C.N. Catton. Inter-observer variability of prostate delineation on cone beam computerised tomography images. *Clinical Oncology*, 21:32–38, 2008.

[WCXM20] N. Wang, C. Chen, Y. Xie, and L. Ma. Brain tumor anomaly detection via latent regularized adversarial network. *arXiv preprint*, arXiv:2007.04734, 2020.

[WD01] L. Wenke and X. Dong. Information-theoretic measures for anomaly detection. *Proceedings 2001 IEEE Symposium on Security and Privacy. S P 2001*, pages 130–143, 2001.

[WGM+11] DF Wulsin, JR Gupta, R. Mani, JA Blanco, and B. Litt. Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *Journal of neural engineering*, 8(3), 2011.

[WKJ18] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.

[WKR+17] R. Wedge, J.M Kanter, S.M Sergio Rubio, I. Perez, and K. Veeramachaneni. Solving the "false positives" problem in fraud prediction. *arXiv preprint*, arXiv:1710.07709, 2017.

[WLA+19] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[WMAG07] Y. Wolff, S. Miron, A. Achiron, and H. Greenspan. Improved csf classification and lesion detection in mr brain images with multiple sclerosis. *Medical Imaging 2007: Image Processing, SPIE*, 6512:920–930, 2007.

[WMCW02] W-K Wong, A.W Moore, G.F Cooper, and M.M Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, pages 217–223, 2002.

[WMCW03] W-K Wong, A.W Moore, G.F Cooper, and M.M Wagner. Bayesian network anomaly pattern detection for disease outbreaks. *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 808–815, 2003.

[WPL+17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R.M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *IEEE conference on computer vision and pattern recognition*, page 2097–2106, 2017.

[WSC20] J. Wolleb, R. Sandkühler, and P. C. Cattin. DeScarGAN: Disease-specific anomaly detection with weak supervision. *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pages 14–24, 2020.

[WT11] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 681–688, 2011.

[WWILT+06] Y. Wua, S.K Warfield, I. I. Leng Tan, W. M. Wells III, D. S. Meier, R. A. van Schijndel, F. Barkhoff, and C.R.G. Guttmann. Automated segmentation of multiple sclerosis lesion subtypes with multichannel mri. *NeuroImage*, 32(3):1205–1215, 2006.

[WZ95] R. J. Williams and D. Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. *Developments in connectionist theory. Backpropagation: Theory, architectures, and applications*, page 433–486, 1995.

[WZL+17] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian. Central focused convolutional neural networks: Developing a data-

driven model for lung nodule segmentation. *Medical image analysis*, 40:172–183, 2017.

[WZT+20] Y. Wang, Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He. Double-uncertainty weighted method for semi-supervised learning. *arXiv preprint*, arXiv:2010.09298, 2020.

[WZW04] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–21, 2004.

[WZX+16] K. Wang, Y. Zhao, Q. Xiong, M. Fan, G. Sun, L. Ma, and T. Liu. Research on healthy anomaly detection model based on deep learning from multiple time-series physiological signals. *Scientific Programming*, 2016.

[XCLT19] Y. Xue, S. Cheng, Y. Li, and L. Tian. Reliable deep-learning-based phase imaging with uncertainty quantification. *Optica*, 6:618–629, 2019.

[XDS+19] C. Xue, Q. Dou, X. Shi, H. Chen, and P. Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pages 1280–1283, 2019.

[XRY+15] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 8.1–8.12, 2015.

[XWLLP06] Y. Xiaoxue, H. Wynne, W.S Lee, and T. Lozano-Pérez. Abnormality detection in retinal images. *Technical Report MIT-3845, MIT, Massachusetts Institute of Technology*, 2006.

[XYLD20] B. Y. Xiaodong, Y. Ying Lv, and T. Denoeux. Evidential deep neural networks for uncertain data classification. *Knowledge Science, Engineering and Management (Proceedings of KSEM 2020)*, pages 427–437, 2020.

[Yan20] X. Yan. Speech network analysis and anomaly detection based on FSS model. *International Journal of Speech Technology*, 2020.

[YC02] D-Y. Yeung and C. Chow. Parzen-window network intrusion detectors. *International Conference on Pattern Recognition*, 4:385–388, 2002.

[YD03] D.-Y. Yeung and Y. Ding. Host-based intrusion detection using dynamic and static behavioral models. *Pattern Recognition*, 36:229–243, 2003.

[YDZ+19a] J. Yang, N. C. Dvornek, F. Zhang, J. Zhuang, J. Chapiro, M. De Lin, and J. S. Duncan. Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[YDZ+19b] J. Yang, N. C. Dvornek, F. Zhang, J. Zhuang, J. Chapiro, M. De Lin, and J. S. Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 11765:255–263, 2019.

[YFH+13] W. Yang, Q. Feng, M. Huang, Z. Lu, and W. Chen. A non-parametric method based on NBNN for automatic detection of liver lesion in CT images. *IEEE 10th International Symposium on Biomedical Imaging*, pages 366–369, 2013.

[YKH01] T. Yairi, Y. Kato, and K. Hori. Fault detection by mining association rules from house-keeping data. *Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2001.

[YLR18] L. Yeo, S. Luewan, and R. Romero. Fetal Intelligent Navigation Echocardiography (FINE) detects 98% of Congenital Heart Disease. *Journal of ultrasound in medicine*, 37(11):2577–2593, 2018.

[YMR16] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[YPGDV19] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint*, arXiv:1906.09686, 2019.

[YTB+17] Y. Yoo, LYW. Tang, T. Brosch, DKB. Li, S. Kolind, I. Vavasour, A. Rauscher, AL. MacKay, A. Traboulsee, and RC. Tam. Deep learning of joint myelin and t1w mri features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. *Neuroimage Clinical*, 17:169–178, 2017.

[YTWM04] K. Yamanishi, Ji. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8:275–300, 2004.

[ZCCK05] J. Zhou, K.L. Chan, V.F.H. Chong, and S.M. Krishnan. Extraction of brain tumor from mr images using one-class support vector machine. *IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6411–6414, 2005.

[ZCLZ16] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, 48:1100–1109, 2016.

[ZCM+99] L. Zheng, A. K. Chan, G. McCord, S. Wu, and J. S. Liu. Detection of cancerous masses for screening mammography using discrete wavelet transform-based multiresolution markov random field. *Journal of Digital Imaging*, 12:18–23, 1999.

[ZCY+17] Y. Zhang, W. Chen, C.K Yeo, C.T Lau, and B. S. Lee. Detecting rumors on online social networks using multi-layer autoencoder. *2017 IEEE Technology & Engineering Management Conference (TEMSCON)*, pages 437–441, 2017.

[ZDW20] L. Zhou, W. Deng, and X. Wu. Unsupervised anomaly localization using vae and beta-vae. *arXiv preprint*, arXiv:2005.10686, 2020.

[ZGC+20] K. Zhou, S. Gao, J. Cheng, Z. Gu, H. Fu, Z. Tu, J. Yang, Y. Zhao, and J. Liu. Sparse-Gan: Sparsity-Constrained Generative Adversarial Network for Anomaly Detection in Retinal OCT Image. *17th IEEE International Symposium on Biomedical Imaging, ISBI 2020, Iowa City, IA, USA, April 3-7, 2020*, pages 1227–1231, 2020.

[ZGL+18] W. Zhang, W. Guo, X. Liu, Y. Liu, J. Zhou, B. Li, Q. Lu, and S. Yang. Lstm-based analysis of industrial iot equipment. *IEEE Access*, 6:23551–23560, 2018.

[ZGMO19] H. Zhang, I. J Goodfellow, D.N Metaxas, and A. Odena. Self-attention generative adversarial networks. *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 97:7354–7363, 2019.

[ZJM+01] Z. Zhang, L. Jun, C. Manikopoulos, J. Jorgenson, and J.L. Ucles. HIDE: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification. *Proceedings of IEEE Workshop on Information Assurance and Security*, pages 85–90, 2001.

[ZKP+18] D. Zimmerer, S A. A. Kohl, J. Petersen, F. Isensee, and K. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint*, arXiv:1812.05941, 2018.

[ZLAL20] M. Z. Zaheer, J.-h Lee, M. Astrid, and S-I Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020.

[ZLCH03] J. Zhou, T-K. Lim, V. Chong, and J. Huang. Segmentation and visualization of nasopharyngeal carcinoma using MRI. *Computers in Biology and Medicine*, 33(5):407–424, 2003.

[ZLS+19] F. Zhang, L. Luo, X. Sun, Z. Zhou, X. Li, Y. Yu, and Y. Wang. Cascaded generative and discriminative learning for microcalcification detection in breast mammograms. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12570–12578, 2019.

[ZLZ+20] R. Zhang, C. Li, J. Zhang, C. Chen, and A. G. Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *International Conference on Learning Representations*, 2020.

[ZPK+21] D. Zimmerer, J. Petersen, G. Köhler, P. Jäger, P. Full, T. Roß, T. Adler, A. Reinke, L. Maier-Hein, and K. Maier-Hein. Medical out-of-distribution analysis challenge. *Zenodo*, 2021.

[ZPP+20] H. Zheng, SMM Perrine, MK Pitirri, K. Kawasaki, C. Wang, J. T. Richtsmeier, and D. Z. Chen. Cartilage segmentation in high-resolution 3d micro-ct images via uncertainty-guided self-training with very sparse annotation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 802–812, 2020.

[ZRM+19] Z. Zhang, A. Romero, M. J. Muckley, P. Vincent, L. Yang, and M. Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2049–2058, 2019.

[ZS15] M. Zareapoor and P. Shamsolmoali. Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48:679–685, 2015.

[ZSGL07] K. Zhang, S. Shi, H. Gao, and J. Li. Unsupervised outlier detection in sensor networks using aggregation tree. *Advanced Data Mining and Application*, pages 158–169, 2007.

[ŻSTI+21] A. Żytkowski, R. Shane Tubbs, J. Iwanaga, E. Clarke, M. Polguj, and G. Wysiadecki. Anatomical normality and variability: Historical perspective and methodological considerations. *Translational Research in Anatomy*, 23:100–105, 2021.

[ZWBC16] J. Zhang, Y. Wu, J. Bai, and F. Chen. Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers. *Transactions of the Institute of Measurement and Control*, 38:435–451, 2016.

[ZWH+18] L-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan, K. Wang, and D. Hu. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine*, 30:74–85, 2018.

[ZXP+20] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, and Y. Xia. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *IEEE Transactions on Medical Imaging*, 2020.

[ZXY+20] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao. Encoding structure-texture relation with P-Net for anomaly detection in retinal images. *Computer Vision - ECCV 2020 - 16th European Conference*, pages 360–377, 2020.

[ZYC+19] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R.X Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2019.

[ZYW+19] P. Zheng, S. Yuan, X. Wu, J. Li, and A. Lu. One-class adversarial nets for fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):1286–1293, 2019.