Context-aware Deep Learning in Medical Image Analysis

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Helen Schneider
aus
Aachen

Bonn 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter\Betreuer:	Prof. Dr. Rafet Sifa
Gutachter:	Prof. Dr. Christian Bauckhage
Tag der Promotion:	24.03.2025
Erscheinungsjahr:	2025

Abstract

Deep Learning (DL) has demonstrated outstanding performance in the area of medical image analysis, with models achieving levels of accuracy that may exceed those of human experts across various applications. This highlights the significant potential of DL-based diagnostic decision support systems to optimize clinical workflows and improve patient outcomes. However, challenges remain that limit the clinical impact and advancements of DL-based systems, such as the need of interpretability or extensive labeled data sets. Since annotation typically requires expert knowledge, generating large-scale annotated data sets presents significant time- and cost constraints. In the following thesis, we address these challenges by incorporating prior contextual information into the DL process, a method known as context-aware DL. Specifically, we focus on two types of contextual information: expert knowledge and prior insights regarding the label quality. The following open research questions are addressed to support the advancement of DL-based decision support systems: 1) Can expert knowledge be leveraged to mitigate current challenges in medical image analysis, and if so, how? 2) Is it possible to utilize contextual information to attain good performance in the multi-label classification of medical image data despite a substantial proportion of missing labels, and if so, how? 3) Is it feasible to integrate contextual information about the label quality to enhance performance when handling data sets with label noise, and if so, how?

Firstly, we integrate expert knowledge regarding the elements of bilateral symmetry of the lung fields into the DL method to automatically detect lung diseases in chest X-ray data. The symmetry-aware architectures and loss function surpass state-of-the-art data-driven DL baselines, enhancing interpretability and data efficiency. To further investigate the potential of expert knowledge as contextual insight, we examine the context-aware analysis of lumbar spine magnetic resonance imaging scans. We demonstrate that contextual information can be integrated through a two-step process. The given work presents an expert knowledge system that utilizes data-driven segmentation masks of the most crucial entities, enabling interpretable diagnostic decision support. These methods emphasize the benefits of context-aware DL based on expert knowledge to address crucial challenges in the medical image analysis field.

Moreover, we focus on the contextual information based on prior insights regarding the label quality to address missing labels or label noise. Proposed novel loss functions achieve high classification performance, surpassing state-of-the-art methods when handling single positive multi-label training for medical images. Additionally, we present a context-aware pre-training strategy to efficiently utilize automatically generated labels, significantly reducing the annotation costs. The context-aware method surpasses purely data-driven training (e.g. not distinguishing between manually and automatically generated labels), highlighting the potential of context-aware training pipelines.
Finally, we introduce a novel context-aware loss function, that achieves remarkable performance even

in the presence of label noise. This given context-aware loss enables the abstaining of potentially noisy samples by integrating insights about the expected label noise as a form of regularization. The presented work underscores the significant potential of context-aware DL to mitigate the adverse effects of missing labels or label noise.

In addition to the introduction of novel context-aware DL methods that address the relevant research questions, we tackle further significant shortcomings in the domain of medical image analysis. We evaluate both proposed and state-of-the-art DL methods within this field, with a particular emphasis on German patient cohorts. This work bridges state-of-the-art DL research and the crucial area of medical image analysis, presenting initial investigation of relevant DL methods for the domain. Consequently, our work enhances the assessment of applicable DL methods for real-life medical image use cases, ultimately improving their potential clinical impact.

Overall, this thesis contributes to the advancement of context-aware DL-based diagnosis decision support systems, aiming to alleviate the workload of medical professionals in their daily practice while enhancing patient outcomes.

Acknowledgements

I would like to express my deep gratitude to everyone who has supported me during my time as a PhD student and in the writing of this thesis.

Firstly, I would like to thank my supervisors, Prof. Dr. Rafet Sifa and Prof. Dr. Christian Bauckhage, for accepting me as their PhD student and for trusting me to pursue my own research topics. I especially want to thank Rafet as my advisor over the last few years. His valuable discussions, critical thinking, and endless motivation greatly benefited this thesis.
Secondly, I would also like to thank all my colleagues and co-authors for the engaging research discussions and their valuable feedback.
Finally, I want to thank my family and friends for their endless support over the years and their helpful reviews of this thesis. I would like to thank my boyfriend Felix, who has supported me over the past years with his kind nature, providing me with the necessary support during the challenging times of my PhD.

Thank you all!

Contents

CHAPTER 1

Introduction

Deep Learning (DL) represents a rapidly advancing domain that is profoundly impacting various industries, through innovations such as driver-assistance systems, translation tools or customer chatbots. In recent years, DL technologies have demonstrated unprecedented exponential growth and will continue to have an even more significant influence on our everyday lives in the future.

In the realm of medical image analysis, DL technologies have demonstrated significant strides in performance, with models showing the potential to achieve diagnostic accuracy that rivals or even surpasses that of human experts across various applications [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Such advances underscore the significant potential of DL-based diagnostic decision support systems (DDSS). Despite these highly relevant achievements, challenges that hinder the clinical impact of DL-based DDSS, persist in the field of DL-based medical image analysis. Key issues include the necessity for extensive labeled training data and the need for enhanced interpretability [4, 9]. The aim of the following thesis is to mitigate these challenges by incorporating prior contextual information into the modeling process, referred to in this thesis as context-aware DL. The area of medical image analysis is well-suited for the application of context-aware DL due to the abundance of available contextual information.

In the following thesis, we demonstrate the significant potential of context-aware DL in the medical imaging field [12, 13, 14, 15, 16, 17, 18]. A *wide variety of novel context-aware methods* is introduced, incorporating two types of contextual information: expert knowledge and prior insights about label quality. Among other innovations, we propose context-aware loss functions that facilitate robust training, even with missing or low-quality ground truth data. Therefore, the context-aware research addresses current challenges in medical image analysis, such as the necessity of extensive labeled data sets with high annotation quality [4].

Moreover, this work connects state-of-the-art DL research and the highly relevant domain of medical imaging. We present *initial investigations* into relevant DL research fields for the complex application field medical image analysis. For instance, we present the first evaluation of single positive multi-label training for medical imaging [15]. We deepen the investigation of relevant DL concepts not only by proposing novel context-aware methods but also by expanding evaluation beyond computer vision benchmark data sets [15, 16, 17, 18]. This enables a more thorough assessment of the applicability of DL techniques to complex real-world issues, such as cancer detection in medical images. Alongside

utilizing public data sets for reproducibility, we use in-house data sets from German healthcare facility for our investigation. We aim to enhance the assessment of context-aware methods for developing efficient DDSS, ultimately improving clinical impact.

Overall, this thesis contributes to the development of context-aware DL-based DDSS, aiming to decrease the burden on medical professionals in their everyday tasks while improving patient outcomes.

1.1 Motivation

In this section, we explore the motivation behind the topic of context-aware DL in medical image analysis and outline the focus of the investigated research.

The analysis of medical image data is highly relevant for clinics and private practices to ensure optimal patient outcomes. Medical images provide a detailed representation of internal body structures, offering valuable information for medical professionals [19]. They are essential for diagnosing a wide range of conditions, from fractures and tumors to cardiovascular and neurological diseases [19, 20, 21, 22, 23]. Early and accurate diagnosis through medical imaging enables timely and effective treatment, significantly improving patient outcomes [19, 24, 25]. Furthermore, medical images are invaluable for treatment planning and monitoring, aiding in the assessment of the effectiveness of implemented treatments [1, 19, 26].
Despite these benefits, the analysis of medical images presents two significant challenges: first, it involves complex data interpretation requiring extensive expertise; second, the sheer volume of daily patient scans creates a time-consuming burden on clinical workflows. These challenges significantly impact the efficiency of both hospitals and private practices, potentially affecting patient care quality and timely diagnoses [3, 27, 28]. Therefore the analysis of medical image data represents a significant cost factor in the healthcare system [29, 30].

Consequently, medical image analysis emerges as a highly relevant field in healthcare technologies, facilitating the automatic detection of pathological structures such as vessel obstructions or tumor lesions [31]. In recent years, DL methods have gained significant relevance in the domain of medical image analysis [32]. In various application areas such as cancer detection, DL models are already achieving high performance, potentially surpassing human experts [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Human error is an inevitable aspect of manual image interpretation, influenced by factors such as fatigue, experience level, and cognitive bias [28, 33]. In contrast, DL algorithms offer a consistent and objective alternative, with the capacity for continuous refinement through integration of new data and real-world clinical feedback [34]. This ongoing improvement ensures that DL methods are up-to-date with the latest medical research and evolving patterns of diseases, thereby enhancing diagnostic capabilities. DL is thus capable of supporting medical professionals in the diagnostic process with DDSS, increasing workflow efficiency and productivity [1, 35, 36]. Automated image analysis can drastically reduce the time required to interpret medical images, enabling physicians to focus on more complex or urgent cases and make faster decisions [33, 36]. Another application of DL models is the triage process in hospitals. By automatically assessing patient conditions based on medical data, DL methods can establish prioritization systems, allowing medical staff to allocate their attention and

resources to the most urgent cases. Overall, DL-based DDSS can optimize the workflow of clinics or private practices, alleviate the workload of medical staff and improve therapeutic outcomes [1, 36]. While initial implementation requires investment, DL methods have the potential to reduce long-term healthcare costs and to enhance patient care quality [36]. These improvements are crucial for maintaining effective healthcare systems, especially given aging populations and the growing shortage of medical professionals [37]. The dual benefits of cost reduction and care enhancement make DL particularly valuable in addressing these pressing healthcare challenges. Further information about medical image analysis and DL-based DDSS is outlined in [1, 31, 32].

Various studies underline the remarkable successes already achieved in the field of data-driven medical image analysis. For instance, in [8], the authors present a single end-to-end trained DL method for the classification of skin lesions. The performance of this data-driven method is evaluated against 21 board-certified dermatologists on biopsy-proven clinical images. The findings demonstrate that the DL model achieves performance on par with all tested medical experts, underscoring that data-driven models can detect skin cancer with a competence level comparable to that of a dermatologist. However, it is important to note that the training data set consisted of 129,450 clinical images labeled by dermatologist. In [3], the authors demonstrate that DL achieves remarkable good performance identifying lung nodules. More than 99% of the lung nodules in chest X-ray scans can be detected by a data-driven network. They use three different data bases, comprising 1,314 nodules. The image data is manually annotated by radiologists.
In [6], the authors investigate the potential of DL methods for automatic chest X-ray interpretation. They focus on multi-label classification, an area particularly pertinent to medical image analysis, as medical images can simultaneously exhibit several disease features. They achieve remarkably performance, outperforming at least 2 of 3 radiologists in the detection of four clinically relevant pathologies. The data set comprises 224,316 chest radiographs of 65,240 patients. Furthermore, the authors of [10] demonstrate that data-driven models surpass medical professionals in detecting breast cancer in mammography. They utilized 170,230 mammograms for training and evaluation. The annotation is biopsy confirmed or based on at least 1-year follow up image data. Recent reserch [38] introduces the foundation model MedSAM , which facilitates universal medical image segmentation, achieving highly accurate segmentation results across various tasks and image modalities. However, the model is based on an extensive medical image data set comprising 1,570,263 image-mask pairs. Please refer to the following publications for further information regarding the exceptional successes of DL-based medical image analysis methods [4, 39, 40].

These studies not only highlight significant achievements in the field of medical image analysis but also underscore the associated challenges. The methods utilized depend heavily on extensive data sets to achieve remarkable performance, typically requiring thousands of samples. To generate high-quality annotations, expert knowledge is often leveraged, revealing the cost and time constraints. Moreover, many studies fail to address the interpretability of the implemented methods, which poses additional challenges. These issues impede the development of efficient DL-based DDSS and their integration into clinical workflows, thus preventing the full potential of DL-based medical image analysis from being realized. In this thesis, we therefore address the following crucial challenges in the medical imaging domain:

Extensive Data Sets The complex DL models required for medical applications often involve

millions of parameters. Consequently, these algorithms demand vast quantities of annotated data to achieve adequate performance [36, 41]. Particularly in medical image analysis, the number of samples included in a data sets is typically small [4]. This issue arises from the high costs associated with data collection. Medical images are obtained from various modalities, including computed tomography, ultrasound imaging and magnetic resonance imaging, all of which entail significant expenses and labor-intensive processes [4]. Furthermore, for use cases involving rare diseases, gathering sufficient image data for the training process presents a significant challenge [4]. Common medical imaging methods often rely on annotated data [4, 36, 42]. The generation of extensive labeled data sets imposes significant time and financial burdens, as expert knowledge is typically necessary for manual annotation, making the process both labor-intensive and costly in the medical field [4]. One consequence of the absence of extensive annotated medical data sets might be that trained DL models are prone to overfitting. While these models achieve exceptionally well accuracy performance on training data, they struggle to generalize when analyzing new data from the problem domain [4].

Annotation Quality Poor label quality is a prevalent issue in many medical image data sets. The associated label noise (i.e. inaccurate annotations) can stem from various factors, including human annotators' limited attention or expertise, the inherently subjective nature of the labeling process, or inaccuracies in automated labeling systems [42, 43]. The inclusion of low-quality annotations can hinder the learning process and diminish the model's generalization capability, as demonstrated by many studies [43, 44, 45]. Various methods have been proposed to successfully handle label noise in data-driven training. However, a wide variety of these methods have been developed and evaluated on large-scale classification data sets [43, 44, 45, 46]. The introduction and evaluation of DL methods to handle label noise for complex medical image use cases therefore represent a highly pertinent research area for the development of efficient DL-based DDSS [43].
Moreover, in a multi-label training setting, each sample can be assigned multiple labels simultaneously [47]. This is especially relevant for medical image analysis, as image modalities can exhibit several disease features concurrently. Missing labels refer to multi-label data sets where not all label information is available for a sample, leaving some labels unobserved [47]. Such missing labels may arise due to time or cost constraints during the annotation process, posing a significant challenge for supervised DL training. Please refer to Section 1.2 for a detailed introduction to label noise and missing labels.

Interpretability Although DL has shown exceptional performance in the medical image analysis domain, the black-box nature of the models presents a significant issue. The implemented methods are opaque, non-intuitive, and difficult for end-users to understand. This lack of interpretability, trust, and transparency represents a challenge for the utilization of DL models to optimize workflows. In critical areas such as healthcare, it is crucial to achieve interpretable decisions to build reliable systems. The lack of interpretability may impede their adoption into clinical settings [9, 26, 36, 48]. A detailed introduction to the topic of interpretability is provided in Section 1.2.

High Performance Given the significant influence of DL-based DDSS on the diagnostic process and subsequent patient outcomes, it is imperative that these models achieve exceptional accuracy, ideally matching or surpassing that of medical experts [34].

As demonstrated in the previous section, medical image analysis often focuses on purely data-driven methodologies. Prior information about the considered use case and utilized data is not integrated into

the modeling process. To overcome these challenges, particularly the high generalization requirements, existing studies often use computer vision techniques such as augmentation or regularization techniques [4, 49, 50, 51, 52]. However, fundamentally, these methods do not integrate new information into the DL methods. Valuable insights can be gained from the contextual information of the considered use case to address current challenges in the medical image analysis domain. Despite the extensive literature on context-aware solution strategies [4, 53, 54, 55, 56, 57, 58, 59, 60, 61], a standardized definition of context-aware DL remains elusive. In this work, we define **context-aware DL** as the integration of additional contextual information in the modeling process leveraging DL. Particularly, the domain of medical imaging lends itself to the application of such techniques due to the abundance of contextual information in place.
We focus on two specific types of contextual information:

1. **Expert Knowledge:** Since human experts require extensive training to analyze complex data manually [4], there is a wealth of prior information available. Expert knowledge encompasses how physicians navigate images, the specific areas they typically focus on, the features they prioritize, and the anatomical prior knowledge they employ. This knowledge is accumulated, summarized, and validated by numerous experts over many years, drawing from a vast array of cases [4]. Dermatologists, for instance, apply established ABCD's rules when examining melanoma [62]. Prior expert knowledge can be integrated into the modeling process to address current challenges in the medical imaging domain, such as data-efficient training [4].

2. **Label Quality:** Contextual information about data sets, particularly label quality, is often available in medical image analysis. We focus on scenarios involving missing labels or label noise, which can impair the model generalization capability [44, 45, 47, 63]. Label quality variations may stem from automated annotation processes or annotators with diverse expertise levels [42, 43]. Identifying which labels were verified by experienced medical experts may provide valuable insights into overall label reliability. This contextual information can be leveraged to enhance DL model accuracy and robustness in medical imaging applications.

The aim of this work is to address crucial challenges in the medical image analysis domain by proposing novel context-aware solution strategies, incorporating valuable insights about the use case and the used data. To achieve this, the following thesis discusses the given research questions:

- Can expert knowledge be leveraged to mitigate current challenges in medical image analysis, and if so, how?
- Is it possible to utilize contextual information to attain good performance in the multi-label classification of medical image data despite a substantial proportion of missing labels, and if so, how?
- Is it feasible to integrate contextual information about the label quality to enhance performance when handling data sets with label noise, and if so, how?

Although the definition of performance can vary in the DL domain (e.g. computational or accuracy performance), this work refers to performance as a measurement how effectively the proposed DL

model solves a given task (e.g. classification of anomalies in medical image data).

In addition to the previously discussed challenges in current medical image analysis research, there are further relevant shortcomings that this thesis addresses.
There exists a gap between state-of-the-art DL research and DL-based analysis of medical image data, as novel methods are frequently evaluated on benchmark computer vision data sets such as CIFAR10 or MS-COCO [64, 65, 66, 67, 68, 69]. Consequently, there is limited insight into the potential of these methods for the analysis of complex real-life use cases, such as medical image analysis. Furthermore, the evaluation of the state-of-the-art DL methods on medical image data sets, if present, is typically based on public benchmark data sets. These often only partially reflect a realistic clinical setup for particular medical conditions. As a result of the introduced biases, DL research in the medical application field rarely achieves a significant clinical impact [70]. Additionally, due to the stringent data protection regulations in Germany, public data sets are frequently provided by foreign research facilities [71, 72, 73, 74]. The highly relevant evaluation on routine data from German practices and clinics is not sufficiently addressed. While DL-focused clinical journals explore the analysis of clinical in-house data sets, they often implement established DL methods rather than proposing novel algorithms [33, 75, 76, 77]. This thesis aims to bridge the gap between current state-of-the-art DL research and the complex application field of medical image analysis, thereby supporting the development of more efficient context-aware DL-based DDSS with enhanced clinical impact.

In summary, the aim of this work is to propose context-aware DL methods specifically tailored for medical image analysis, addressing current challenges within the application domain. The discussed research introduces novel context-aware solutions, with evaluations conducted, among others, on routine data from German healthcare facilities. This represents a valuable contribution for researchers, clinicians and companies interested in the development of efficient DL-based DDSS.

1.2 Technical Background

To maintain the self-contained nature of this thesis, the following section provides essential technical background on Machine Learning (ML) and neural networks. We focus on the training processes of neural networks, exploring key architectures and loss functions.

1.2.1 Machine Learning

ML algorithms fundamentally aim to extract patterns from observed data to make predictions or decisions about new, unseen data for the same task. We can structure the modeling process in mainly two different phases: training and testing. A model is trained on observed data and tested on unseen test data after training. The aim is to perform well on both sets, although the test data is not considered during training. This behavior is referred to as **generalization** [78]. This generalization capability is crucial for the model's real-world applicability and effectiveness.

Different training scenarios can be broadly categorized into supervised, unsupervised, and reinforcement learning based on the annotation and feedback of the data [79]. This thesis focuses on supervised learning, where each data sample consists of an input ($\boldsymbol{x}$) and output ($\boldsymbol{y}$) pair. The objective is to design a model that can map inputs to the corresponding output. These output values are called targets,

ground truth, or labels [78, 79]. Supervised learning problems can be categorized into two main tasks, classification and regression. The former assigns a category to each sample, such as classifying patient conditions, while the latter predicts a real value for each item, for instance modeling a patient's recovery time [16, 78, 80]. In this work, we focus on classification problems. For these tasks, the original string targets are converted into vectors of the size of the number of classes, C. The typically binary target variables $\boldsymbol{y}_i \in \{0, 1\}$ indicate the given classes [79].

The aim of ML algorithms is to find the optimal function approximation $f^* \in \mathcal{F}$, also known as the best **hypothesis** within the hypothesis set $\mathcal{F}$, of an unknown relationship $l : \mathcal{X} \to \mathcal{Y}$ between inputs $\boldsymbol{x} \in \mathcal{X}$ and outputs $\boldsymbol{y} \in \mathcal{Y}$. Since the prior probability distribution over all samples is unknown in practice, we cannot minimize the expected generalized loss to define the best hypothesis f^*. However, this can be estimated by the **empirical risk** $R_{\mathcal{D}}$, which is calculated based on a given data set $\mathcal{D} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1,\dots,N}$ with N input-output samples. The estimated optimal hypothesis $\hat{f}^*$ represents the minimum of the empirical risk

$$\begin{aligned} \hat{f}^* &:= \underset{f \in \mathcal{F}}{argmin}\, R_{\mathcal{D}}(f) \\ R_{\mathcal{D}}(f) &= \frac{1}{N} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} L\Big(\boldsymbol{y}, f(\boldsymbol{x})\Big) \end{aligned} \tag{1.1}$$

where L represents a given loss function [81, 82]. In ML research, we often refer to the final function as the model. For reasons of clarity, the ML model is referred to as f in this thesis. Further information about empirical risk minimization is outlined in [81, 83].

The choice of the implemented model for the function class $\mathcal{F}$ depends on the given problem and data. Popular models are decision trees [84], support vector machines [85] and neural networks [82, 86]. Generally, the choice of a model involves a trade-off between a more complex model that fits the training data well, and a simpler model that may generalize better on unseen data [82]. This effect is called **bias-variance-trade-off**. A complex function for a relatively small sample size allows a strong variance. This can lead to **overfitting** of the training data and poor generalization on unseen data [87]. Conversely, a function that is too simple for an extensive sample size and complex problem set up can create a high bias, leading to **underfitting** and insufficient performance [87]. The aim of the model selection is to neither under- nor overfit [78, 79, 87, 88]. Regularization techniques, such as L^2 parameter regularization, can be employed to avoid or mitigate the effects of overfitting. A more detailed overview of regularization methods is outlined in [79, 82].
A variety of ML algorithms rely on hyperparameters to regulate their trainings behavior. These hyperparameters are not adjusted by the learning algorithms, as determining their specific values from the training set is often inappropriate (e.g. regularization hyperparameters) [87]. To address this, we utilize a validation set to estimate the generalization error during training. It's important to note that test, validation and training samples are strictly separated [87].

While traditional ML algorithms such as support vector machines achieve remarkable performance across a range of relevant problems, they face challenges in key applications such as speech and object recognition [87]. One significant limitation of these traditional ML methods is their difficulty in

handling high-dimensional data, since many ML problems become extremely complex when dealing with high-dimensional data [87]. This issue is referred to as the **curse of dimensionality**. A key concern is that the number of possible unique configurations of a variable set grows exponentially with an increase in the number of variables. A detailed introduction in the curse of dimensionality is outlined in [79]. Conventional ML techniques are limited in their ability to process natural data in its raw form, such as pixel values. Careful engineering is often required to design an efficient feature extractor, transforming raw data into suitable representations for final classification [89]. A more detailed discussion of the challenges posed by conventional ML techniques is outlined in [87]. The evolution of deep neural networks specifically targets the limitations of traditional machine learning approaches, achieving remarkable generalization performance across a wide range of complex problems, particularly those involving high-dimensional data. Given our focus on medical image analysis, we will concentrate on deep neural networks, exploring their architectures, training methodologies, and applications in healthcare.

1.2.2 Neural Networks

Neural networks are complex non-linear ML models. They commonly comprise multiple layers, including one input, potentially several hidden, and one output layer. These layers represent a composition of neurons or units, which are linked by weighted connections . The input of a single unit represents the weighted sum of the outputs of connected units. This value is then processed by a **non-linear activation function** to calculate the output of the given unit. Among others, popular non-linear activation functions include the rectified linear unit (ReLU), defined as $h(z) = max(0, z)$, and the sigmoid function, defined as $h(z) = 1/(1 + e^{-z})$. A multilayer network is a concatenation of the individual layers and their activation functions, as presented for a 2-hidden-layer network

$$f(\boldsymbol{x}) = h_3(\boldsymbol{W}_3 \cdot h_2(\boldsymbol{W}_2 \cdot h_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3), \tag{1.2}$$

where $\boldsymbol{x}$ represents an input vector, h_i are the non-linear activation functions and $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are the weight matrices and biases. Note that the weights and biases represent the parameters that are tuned during the training of the neural network to obtain a good mapping function. In the following, the overall trainable parameters of the network are represented as $\boldsymbol{\theta}$. Figure 1.1 represents a **feedforward neural network**, all connections are, as the name suggests, in one direction. Typically each unit is connected to all units of the preceding layer. This is referred to as a fully connected layer [79, 87].

Neural networks are trained to determine appropriate connection weights, with the objective of achieving high generalization capability. The training process starts with forward propagation, where input data is passed through the network to generate an output. The error between the prediction and ground truth is calculated using a loss function. The backpropagation algorithm allows the calculated loss information to flow backwards through the network, enabling the computation of the gradients of the loss function with respect to the model parameters [79, 90].
An optimization algorithm, such as Adam [91], use these gradients to update the weights in the network. Most common optimizers are based on the fundamental **Gradient Descent** method, which updates the model parameters in the direction of the negative gradients to minimize the loss [79]. Note that each update step requires the entire training data set to be processed.

Stochastic Gradient Descent is a extension of gradient descent that optimizes model parameters by

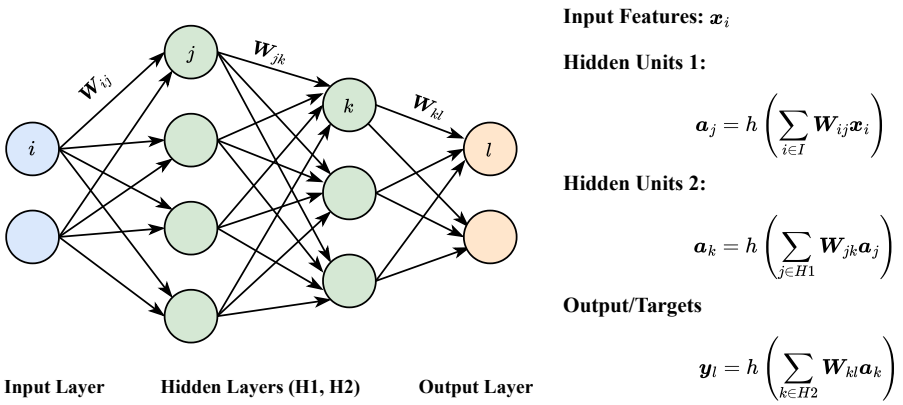


Figure 1.1: The given figure represents a simple **feedforward neural network**, inspired by [89]. The model has 1 input, 2 hidden, and 1 output layer. The input features are represented by $\boldsymbol{x_i}$, $\boldsymbol{y_l}$ represents the outputs. The values at each unit are calculated during a forward pass as a weighted sum of the previous layer, followed by the application of a non-linear activation function h. The model weights represent the parameters that are tuned during the training process. Please note that all connections are weighted. For clarity, we neglected the bias parameter in the figure.

computing gradients on randomly sampled subsets of the data, known as mini-batches. This approach offers several advantages, including reduced computational cost and potential regularization effects [79]. This cycle of forward propagation of the entire data set, loss calculation, backward propagation, and weight update is repeated over multiple iterations, known as epochs.

Through these iterative adjustments, neural networks enhance their ability to make accurate predictions, ultimately achieving the desired performance for the given task.

Deep Learning (DL) involves the training of neural networks with several hidden layers, leading to more complex model architectures [87, 92]. As DL enables efficient representation learning of data with multiple levels of abstraction, these methods have remarkably improved the state-of-the-art for complex use cases such as speech and object recognition. Given that the following work focuses on the analysis of raw medical imaging data, DL represents a key aspect of state-of-the-art data-driven research [89]. Due to the increased number of parameters, extensive data sets are required for effective model training to ensure good generalization performance [41, 93, 94]. As discussed in the previous Section 1.1, this marks a particular challenge in the medical imaging domain, where the generation of large-scale annotated data sets poses significant cost and time constraints.

The initialization of the given network parameters can significantly influence the optimization process and the generalization ability of the model. Parameters are often initialized randomly and follow a uniform or Gaussian distribution [95]. Alternatively, neural network parameters can also be initialized by reusing parameters from a pre-trained model, which can then be fine-tuned on the given training data for the target task. This approach can not only enhance the optimization process but also lead to better generalization and enable data-efficient training. In recent years, supervised pre-training on the ImageNet data set has become a common **transfer learning** approach for computer vision tasks [93, 96]. Transfer learning facilitates the integration of prior knowledge regarding the pre-trained task

into the DL training process [4].

Transfer learning can be leveraged for various ML (classification) tasks to enhance the model's capability to generalize. In the realm of DL, two significant types of classification tasks are multi-class and multi-label training. Multi-class training is used when each data instance belongs to one and only one of several possible classes [79]. Conversely, multi-label training is employed when each data instance can belong to multiple classes simultaneously. For instance, a CT scan can exhibit features of several diseases [63]. The implemented loss function depends on the classification task. The **cross-entropy** loss (CE) is a popular loss function for multi-class classification

$$\begin{aligned} L_{\text{CE}}\Big(f(\boldsymbol{x}), \boldsymbol{y}\Big) &= -\sum_{i=1}^{C} \boldsymbol{y}_i \log\Big(f_i(\boldsymbol{x})\Big) \\ \frac{\delta L_{\text{CE}}\Big(f(\boldsymbol{x}), \boldsymbol{y}\Big)}{\delta \boldsymbol{\theta}} &= -\sum_{i=1}^{C} \boldsymbol{y}_i \frac{1}{f_i(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x}) \end{aligned} \tag{1.3}$$

where $\boldsymbol{\theta}$ is the set of parameters of the model f, $\boldsymbol{y}_j$ corresponds to the j'th element of the label vector of the sample $\boldsymbol{x}$ and C represents the number of considered classes [44].
For multi-label classification, the common **binary cross-entropy** loss (BCE) [63] is defined as

$$\begin{aligned} L_{\text{BCE}}\Big(f(\boldsymbol{x}), \boldsymbol{y}\Big) &= -\frac{1}{C}\sum_{i=1}^{C} \Big[\boldsymbol{y}_i \log\Big(f_i(\boldsymbol{x})\Big) + (1-\boldsymbol{y}_i)\log\Big(1-f_i(\boldsymbol{x})\Big)\Big] \\ \frac{\delta L_{\text{BCE}}\Big(f(\boldsymbol{x}), \boldsymbol{y}\Big)}{\delta \boldsymbol{\theta}} &= -\frac{1}{C}\sum_{i=1}^{C} \Big[\boldsymbol{y}_i \frac{1}{f_i(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x}) - (1-\boldsymbol{y}_i)\frac{1}{1-f_i(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x})\Big]. \end{aligned} \tag{1.4}$$

The model outputs are commonly interpreted as probability [63]. We achieve this for multi-class classification by applying a **softmax** layer [79]. The k'th element of the softmax function is defined as

$$\text{softmax}(\boldsymbol{o})_k = \frac{e^{\boldsymbol{o}_k}}{\sum_{i=1}^{d} e^{\boldsymbol{o}_i}} \tag{1.5}$$

where $\boldsymbol{o}$ represents an input vector with length d. For each sample $\boldsymbol{x}$, $f(\boldsymbol{x})$ computes its probability of each class in $c \in \{1, \cdots, C\}$ and $\sum_{i=1}^{C} f_i(\boldsymbol{x}) = 1$ [79]. For multi-label classification, several labels can be true simultaneously, therefore, the sigmoid function is often used for each neuron to interpret the outputs as probabilities [97].

Despite great success in various applications, these conventional loss functions are not always suitable for the given problem setup, particularly when noise robustness is a critical factor [44].
In the following thesis, noise robustness refers to a model's ability to perform well in the presence of mislabeled data. In this context, we will refer to an incorrect label as "noisy". Noisy labels can have a strong negative impact on the performance of CE training, because the CE loss function is not inherently **noise-robust** [44]. The implicit weighting factor $1/f_i(\boldsymbol{x})$ in the calculation of the gradients leads to a stronger training influence from samples with high deviation between the prediction and the available ground truth. While this is desirable for a data set with clean labels, it leads to a

strong focus on incorrect information in the case of label noise. Consequently, training with the CE loss may lead to noise overfitting, thereby weakening the generalization performance of the final model [44]. The same applies to the BCE loss. The development of noise-robust classification methods therefore commonly represents a trade-off between noise-robustness and performance enhancement.

In addition to the lack of noise-robustness, the unobserved ground truth information poses a significant challenge for supervised DL methods. As discussed in Section 1.1, it can be prohibitively expensive to generate the large number of annotations required to train deep neural networks. This is especially relevant for multi-label annotations. Given the impracticality of exhaustively annotating every image for all potential classes, a natural trade-off arises between how many images are annotated and how many labels we consider during the annotation process [63]. In practice, not all label information is always available for a sample. This is referred to as **missing** or partially observed labels [47]. For instance, when using clinical routine data, specifically medical reports associated with an image, to annotate chest X-ray scans, it is possible that specific findings are neither negated nor confirmed, resulting in a missing label. **Single positive multi-label** (SPML) training represents one of the most severe cases of partially observed labels [63]. For each training instance, only one positive label is provided, despite the possibility of multiple correct labels. Other labels for a sample are unobserved [63]. SPML training is especially relevant for applications where it is difficult or costly to annotate all relevant labels for each instance, such as medical image analysis [52, 63, 69]. It facilitates the generation of extensive (partially annotated) data sets, with reasonable annotation costs.

The limitations of standard CE\BCE training underscore the critical importance of selecting an appropriate loss function within the training pipeline [98]. Consequently, the development of tailored loss functions that account for specific problem characteristics and data set properties has emerged as a fundamental research area in ML.

Convolutional Neural Networks

Despite their great success, conventional feedforward neural networks have reached their limitations for processing data with spatial or temporal information [79, 87]. As a result, more complex network architectures, such as recurrent neural networks [99] or **convolutional neural networks** (CNNs) have been developed [100]. The first serve the purpose of processing data with temporal information, while the second are designed to process grid data like images [87]. A key characteristic of image analysis is that nearby pixels exhibit stronger correlations than distant ones [79, 82]. Several modern computer vision methods exploit this property to extract **local features** that consider only a small subregion of the input [79, 82]. Local features can be combined in later processing stages to identify high-order features, ultimately facilitating the classification of the image as a whole [79]. In addition, local features from a subregion of the image are likely to be useful in other regions of the image as well [79]. These concepts are integrated into the CNN architecture based on three different properties: local receptive fields, shared weights, and pooling [79, 87]. The general structure of a CNN (used for classification) is visualized in Figure 1.2.

CNNs exploit a composition of two layers: convolutional and pooling layers. The units in the convolutional layer are organized as planes, referred to as **feature maps** [79]. Units of a feature map only consider a small region of the image (or spatially organized units in a subsequent layer). Input

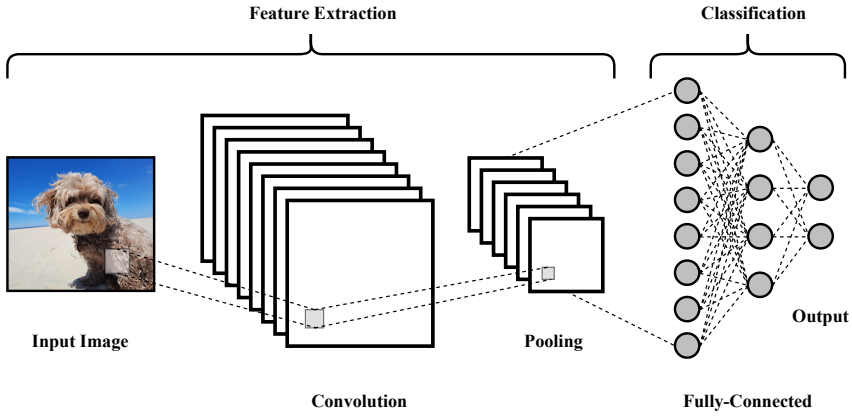


Figure 1.2: The given figure visualizes an example of a CNN architecture for classification. During the first part of the network, we extract important features of the input based on convolution and pooling layers. The applied kernels are visualized as squares. This is followed by fully connected layers for the final classification.

values of this patch are linearly combined using weight and bias parameters and transformed by a non-linear activation function [79]. The units in one feature map all share the same weights [79, 82]. In terms of feature detection, all units in one feature map detect the same pattern but at different locations [79]. Since it is typically necessary to detect several features, a convolutional layer often contains several feature maps. Each feature map has its own set of weight and bias parameters. The applied pattern of parameters is referred to as **kernel** [82]. The application of the kernel onto the pixels of the input image (or spatially organized units in a subsequent layer) is called **convolution** [79, 82]. Please note that the commonly used definition of convolution in DL does not precisely align with the definition in other fields, such as signal processing [82]. A shift in the input data results in a corresponding shift in the feature map activations, while preserving their overall structure. This property forms the basis for the (approximate) invariance of the network outputs to translations or local distortions of the input image [79, 100].

The outputs of the convolutional units represent the input for the pooling layers [79]. Convolutional layers extract relevant features from the previous layer, while pooling layers condense information by merging adjacent units into summary statistics, effectively downsampling the layer size [82, 87]. Among others, a popular pooling approach is the max pooling [101], which calculates the maximum value within a rectangular neighborhood [87]. The application of pooling layers contributes to achieving approximate invariance of the representation to minor translations in the input data [87].

Several stages of convolution and pooling are often layered on top of each other [79]. For each plane in the previous subsampling layer, there may be several feature maps in the following convolutional layer. This compensates for the reduction in spatial resolution by increasing the number of features [79]. Since all weights are learned in a data-driven manner, CNN architectures can generate their own feature extractor, enabling the processing of raw high-dimensional image data [87, 100].

CNN architectures have achieved high performance in various image processing tasks, such as the analysis of medical images [5, 6, 8, 39, 87]. To solve these complex problems, modern architectures

such as Densely Connected Convolutional Networks (**DenseNet**) and Residual Neural Networks (ResNet) are based on millions of parameters [102, 103]. This work leverages, among others, the DenseNet architecture, which is designed to improve the gradient flow and reduce computational complexity [102]. Unlike traditional architectures, DenseNet introduces a dense connectivity pattern. Each layer receives the feature maps from all preceding layers and transfers its own feature maps to all subsequent layers [102]. The architecture substantially reduces the number of parameters, making the network more efficient and easier to train. DenseNet models are characterized by dense blocks and transition layers [102].

In addition to the image classification task, CNN models can also be used for **segmentation** [104]. In order to recognize different objects or regions in the images, each pixel receives a label. It is therefore a pixel-by-pixel classification [104]. Common architectures include U-Net models, introduced in [104], which are based on an encoder-decoder structure. The encoder uses convolution and pooling layers to downsample the image data, capturing the most important features. The decoder upsamples the encoded feature representations to reconstruct the output image [104]. In order to achieve high accuracy segmentations, the decoder combines the corresponding high resolutional features of the encoder with the upsampled outputs [104].

Complex CNN architectures represent black box models, their decision-making process is usually not interpretable [26]. **Interpretability** refers to the ability to understand and explain how a model makes its decisions [48, 82]. It is crucial for ensuring transparency, trust, and accountability, especially in critical applications like medical image analysis [48]. We can use interpretability to enhance the trustworthiness of DL models [48]. Several interpretable methods have been developed for CNNs, such as the Gradient-weighted Class Activation Mapping (**Grad-CAM**) method [105]. This method uses the gradients of a target class flowing into the final convolutional layer to produce a coarse localization map highlighting the relevant areas in the input image for predicting that class [105]. Various extensions, such as Grad-Cam++ [106] are available. In this work we define interpretability and explainability as synonymous, both referring to achieving a better understanding of the model and its behavior.

A further use case for CNNs involves implementing similarity learning tasks based on image data [107]. In this thesis, we aim to analyze two images as input, with the objective of determining whether the pair of data is dissimilar. A widely recognized architecture for this purpose is the **Siamese network** [107, 108]. Two given inputs are processed simultaneously by two CNNs with shared weights [107, 108]. A prevalent loss function used in the training of Siamese networks is the **contrastive loss** [109], defined as

$$L_{\text{contrastive}}(t, D) = t\frac{1}{2}D^2 + (1-t)\frac{1}{2}\Big\{\max(0, m - D)\Big\}^2, \tag{1.6}$$

where t denotes the similarity target ($t = 0$ if the pair is dissimilar), and D represents a distance function, such as the euclidean distance, based on the calculated embedding of the data pair. The parameter m is the leveraged margin, defining the maximum distance up to which dissimilar pairs still influence the loss function. The aim of contrastive loss is to draw similar instances closer in the feature space and push dissimilar instances apart [107, 109]. Siamese neural networks are utilized in medical image analysis, for instance, for the unsupervised detection of outliers [110]. In this thesis, Siamese neural networks are used to integrate expert knowledge about elements of bilateral symmetry

of the human body into data-driven modeling. More detailed information about the used Siamese neural networks in the following work is outlined in the Chapters 2 and 3.

1.3 Applications and Transferability in Medical Image Analysis

In this thesis, we consider various challenges that arise in the processing of medical image data. The following section provides an overview of the clinical use cases and corresponding data. Please note that while this research focuses on X-ray and magnetic resonance imaging scans, the proposed context-aware solutions are *adaptable across a broad spectrum of image modalities and use cases*. Challenges such as label noise, integration of automatically generated annotations or missing labels concern a wide variety of medical use cases, such as skin cancer detection utilizing dermatoscopic images [111], COVID-19 diagnosis utilizing computed tomography scans [112] or surgical tool detection in laparoscopic videos [113]. This underlines the significant impact of our work for the medical image analysis domain, independently of the considered use cases.

1.3.1 Chest X-ray Scans

Chest X-ray (CXR) images serve as an essential diagnostic tool in the everyday workflow of clinics. Due to their time efficiency, cost-effectiveness, and relatively low radiation exposure compared to other imaging modalities CXR scans are extensively used in clinical practice. The resulting high-resolution, 2D grayscale images capture the internal body structure of the thorax, aiding the diagnosis process of many pathological findings [114]. Typically, the images are stored in the Digital Imaging and Communications in Medicine (DICOM) format, ensuring standardized viewing and analysis across different healthcare systems [35, 115]. Additionally, metadata such as patient information is included alongside the image data [115].

Various types of CXR scans offer different insights. Frontal views, visualized in Figure 1.3, provide a detailed overview among others of the hilar structures, heart, lung parenchyma, thoracic wall and mediastinum [114, 116]. In the following sections we focus on processing frontal scans. Posteroanterior CXR images represent the standard view, captured with the patient standing and facing the X-ray film [114]. The anteroposterior CXR scan is often leveraged when the patient is on bed rest, commonly in Intensive Care Unit (ICU) settings [117]. The X-ray machine is placed in front of the patient, while the film is positioned behind them. Compared to the posteroanterior view, the thorax structures are not as clearly visualized and structures far away from the image film, like the heart, appear enlarged due to the direction of X-ray penetration [114, 118].

In this thesis, we use CXR images to detect anomalies, among others, focusing on pulmonary infiltrates, fractures, pleural effusion, pulmonary congestion, pneumothorax, atelectasis, cardiomegaly, pneumonia and edema. Additionally, we detect the misplacement of the central venous catheter.
The automatic analysis of these anomalies can be leveraged as DDSS, among others, to support physicians during their decision or to develop prioritization systems. Further information about CXR data is outlined in [114].

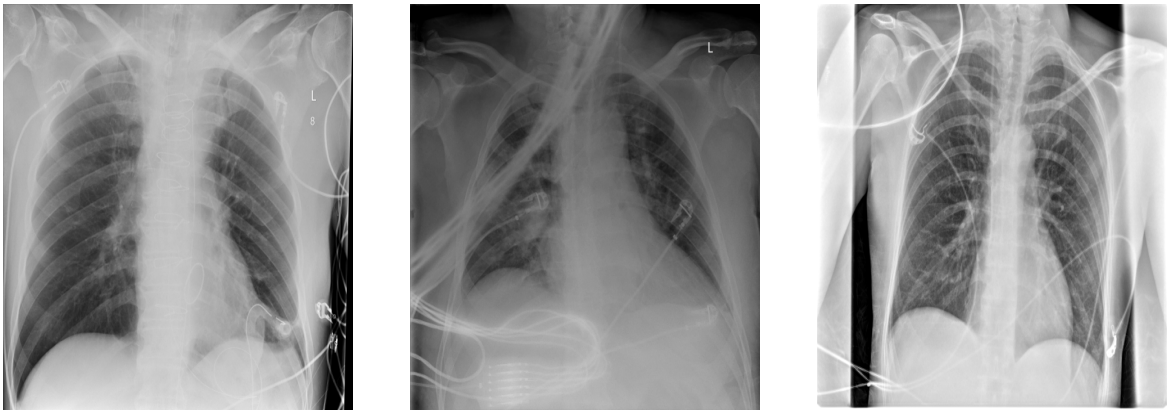


Figure 1.3: Samples of anteroposterior CXR scans provided by the University Hospital Bonn.

1.3.2 Lumbar Spine Magnetic Resonance Imaging Scans

Lumbar spine magnetic resonance imaging (MRI) scans are relevant diagnostic tools employed to visualize the internal structure of the spine, including vertebrae, intervertebral discs, spinal canal and surrounding soft tissues [119, 120, 121]. Lower back pain is ubiquitous in modern society, representing one of the most frequent reasons for consulting a primary healthcare provider [119, 122]. When conservative treatment is ineffective or concerning clinical findings arise, MRI becomes the preferred imaging technique to evaluate the causes of spine pain. Various potential aetiologies with similar clinical presentation exist, such as degenerative changes, infection, and pathological fractures. MRI facilitates the differentiation of these sources of spine pain, enabling appropriate therapy [119]. Typically T1-weighted and T2-weighted scans are produced [121]. The former provides a clear overview of anatomical structure, for instance by visualizing the thecal sac and epidural space more distinctly [121]. The latter highlights fluid containing structures, aiding the detection of anomalies such as disc desiccation [121]. In this thesis we are exploiting lateral scans to detect anomalies. Figure 1.4 therefore visualizes two 2D slices for lateral scans.

The analysis of lumbar spine MRI scans is particularly challenging, a significant variability in the interrater and intrarate agreements is typically observed [123]. There exists no single established grading system for many findings, Lumbar spine MRI scans are often investigated by eye [123, 124]. Additionally, specific distance measurement, such as the height of the vertebral bodies, are used to identify anomalies. The manual measurements represent a time consuming and repetitive task [14, 124]. The automatic analysis of lumbar spine MRI scans can alleviate the workload of medical professionals and improve patient outcome, e.g. by reducing the manual work load and standardizing measurements. Please refer to [119] for further information about lumbar spine MRI scans.

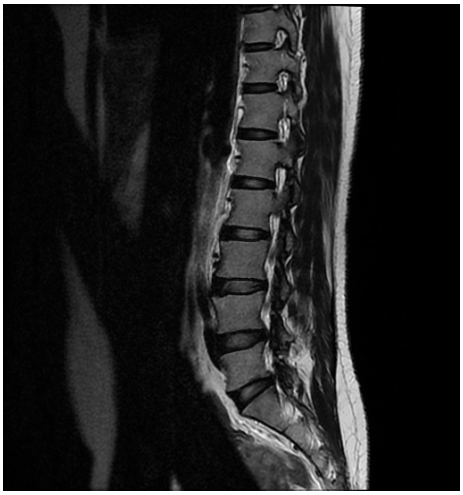



Figure 1.4: Samples of lateral T2-weighted lumbar spine MRI scans, provided by the Evidia GmbH.

1.4 Thesis Contributions

1.4.1 List of Publications

This cumulative dissertation comprises the following peer-reviewed publications, in which I served as the primary contributor. The list is organized by topic and aligns with the order of presentation in this dissertation.

- H. Schneider, M. Lübbering, R. Kador, M. Broß, P. Priya, D. Biesner, B. Wulff, T. Bell Felix de Oliveira, Y.C. Layer, U. I. Attenberger, R. Sifa. **Towards Symmetry-aware Pneumonia Detection on Chest X-rays**. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022.

- H. Schneider, E.C. Yildiz, D. Biesner, Y.C. Layer, B. Wulff, S. Nowak, M. Theis, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Symmetry-aware Siamese Network: Exploiting Pathological Asymmetry for Chest X-ray Analysis**. In *International Conference on Artificial Neural Networks (ICANN)*, 2023.

  The research was honored by the **Springer&ENNS Best Paper Award of the ICANN conference 2023**.

- H. Schneider, D. Biesner, A. Ashokan, M. Broß, R. Kador, S. Halscheidt, B. Bagyó, P. Dankerl, H. Ragab, J. Yamamura, C. Labisch, R. Sifa. **Segmentation and Analysis of Lumbar Spine MRI Scans for Vertebral Body Measurements**. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2023.

- H. Schneider, P. Priya, D. Biesner, R. Kador, Y.C. Layer, M. Theis, S. Nowak, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Is One Label All You Need? Single Positive Multi-label Training in Medical Image Analysis**. In *IEEE International Conference on Big Data (BigData)*, 2023.

- H. Schneider, D. Biesner, S. Nowak, Y.C. Layer, M. Theis, W. Block, B. Wulff, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Improving Intensive Care Chest X-Ray Classification by Transfer Learning and Automatic Label Generation**. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2022.

- S. Nowak, H. Schneider, Y.C. Layer, M. Theis, D. Biesner, W. Block, B. Wulff, U.I. Attenberger, R. Sifa, A.M. Sprinkart. **Development of Image-based Decision Support Systems utilizing Information Extracted from Radiological Free-text Report Databases with Text-based Transformers**. In *European Radiology*, 2024.

  Please note that this represents a shared first authorship.

- H. Schneider, S. Nowak, A. Parikh, Y.C. Layer, M. Theis, W. Block, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems**. Accepted at *International Conference on Neural Information Processing (ICONIP)*, 2024.

  Preprint:
  H. Schneider, S. Nowak, A. Parikh, Y.C. Layer, M. Theis, W. Block, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems**. In *arXiv preprint arXiv:2410.21014*, 2024.

1.4.2 List of Key Contributions

The proposition and evaluation of context-aware DL methods for the analysis of medical image data sets to mitigate current challenges and shortcomings in the field are discussed in the relevant publications. The key contributions are:

Towards Symmetry-aware Pneumonia Detection on Chest X-rays

- Proposition of novel Siamese network architectures for the analysis of CXR data: integration of expert knowledge about the elements of bilateral symmetry of the lung fields in CXR scans
- To the best of our knowledge, the first investigation of symmetry-aware DL architecture for the analysis of CXR scans
- Enhanced performance, data efficiency and interpretability

Symmetry-aware Siamese Network: Exploiting Pathological Asymmetry for Chest X-ray Analysis

- Further investigation of novel symmetry-aware CXR analysis: integration of the expert knowledge about the elements of bilateral symmetry of the lung fields in CXR scans in extended contrastive loss function
- To the best of our knowledge, the first investigation of symmetry-aware loss functions for CXR scan analysis
- Enhanced performance and interpretability

Segmentation and Analysis of Lumbar Spine MRI Scans for Vertebral Body Measurement

- Introduction of an interpretable context-aware solution approach: combination of rule-based expert knowledge system and DL-based segmentation
- Experiments based on a German in-house data set provided by the Evidia GmbH
- Development of interpretable anomaly detection method

Is One Label All You Need? Single Positive Multi-label Training in Medical Image Analysis

- To the best of our knowledge, the first investigation of state-of-the-art SPML training for medical image analysis
- Proposition of novel context-aware SPML loss functions: integrating contextual information about missing labels
- Enhanced training robustness when challenged with missing labels

Improving Intensive Care Chest X-ray Classification by Transfer Learning and Automatic Label Generation

- Investigation of the potential for report-based generated labels with a rule-based natural language processing (NLP) labeler to enable training with extensive annotated data sets
- Proposition of context-aware pre-training strategy to process automatically generated labels
- Initial processing of an in-house ICU CXR data set based on clinical routine data provided by the University Hospital Bonn
- Enhanced performance by leveraging automatically generated annotations

Development of Image-based Decision Support Systems utilizing Information Extracted from Radiological free-text Report Databases with Text-based Transformers

- Exploration of the potential and limitations of using transformer-based report annotations for the development of context-aware image-based DDSS utilizing clinical routine data, provided by the University Hospital Bonn
- Proposition on context-aware pre-training strategy to process automatically generated labels
- Enhanced performance with automatically generated labels and context-aware training

Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems

- Introduction of novel context-aware loss function for the processing of data sets with label noise
- Performance evaluation based on noise simulation using public data and clinical in-house data set with automatically generated real-life noisy labels
- Facilitating noise-robust training for medical image analysis

1.4.3 Contributions Summary

In the following sections, we briefly present the most significant contributions of each chapter and outline the structure of the thesis, particularly the relationships between the respective chapters. Our focus lies in incorporating contextual information into the modeling process to address current challenges in the medical image analysis domain. We introduce a *diverse range of novel context-aware DL methods*. Specifically, our research leverages German clinical routine data, among other approaches, to underscore the practical relevance of our findings in healthcare settings. This work aims to bridge the gap between cutting-edge DL research and the critical field of medical image analysis.

Towards Symmetry-aware Pneumonia Detection on Chest X-rays

A prevalent DL approach for detecting lung disease in CXR scans involves using deep CNNs trained in a data-driven supervised manner. While these methods have shown remarkable performance across various applications [6, 125, 126], challenges such as model interpretability and data-efficient training continue to exist. By incorporating expert knowledge into the data-driven training process, we aim to address these challenges. Since medical experts tend to intuitively compare the left and right lung fields of CXR scans to detect anomalies [118], we leverage the elements of bilateral symmetry of lung fields in CXR scans as contextual information. While the components of bilateral symmetry of the human body have proven to be valuable expert knowledge for various DL methods and medical use cases, their potential for CXR analysis remains underexplored [127, 128, 129].

To the best of our knowledge, this research represents the *initial investigation of the potential of the elements of the bilateral symmetry of the lung fields in CXR scans*, initiating a highly pertinent research area. We propose novel context-aware Siamese networks to integrate the contextual information about the elements of bilateral symmetry of the lung fields in CXR scans. Our findings indicate that context-aware DL (based on the expert knowledge) not only improves performance but also enhances data efficiency and interpretability, both of which are highly relevant key contributions to the medical image analysis domain. Additional results are outlined in Chapter 2.

Central research question: Can expert knowledge (here: elements of bilateral symmetry of the lung fields in CXR scans) be leveraged to mitigate current challenges in medical image analysis, and if so, how?

Symmetry-aware Siamese Network: Exploiting Pathological Asymmetry for Chest X-ray Analysis

Building on the initial positive research results regarding the prior knowledge of the elements of bilateral symmetry of lung fields in CXR scans for context-aware DL methods, we further investigate this contextual information to address current challenges. We introduce novel context-aware DL methods for analyzing CXR data, aiming to enhance the impact of contextual information within the DL pipeline. In addition to adapting context-aware networks, motivated by the previous chapter, we integrate expert knowledge through a novel context-aware contrastive loss function. Based on the integration of the contextual information regarding the elements of pathological bilateral asymmetry of the lung fields in CXR scans, we strive to achieve a stronger differentiation

between pathological (e.g., relevant for the classification task) and non-pathological asymmetries. As in the previous chapter, the proposed context-aware solution represents an end-to-end training approach.

To the best of our knowledge, this research represents the *first investigation of symmetry-aware loss functions for analyzing CXR scans* regarding lung diseases. We achieve improved performance and interpretability, facilitating the differentiation between pathological and non-pathological asymmetries. A comprehensive introduction of the proposed architectures, loss functions, and results is provided in Chapter 3.
The research received the **Springer&ENNS Best Paper Award of the ICANN conference 2023**.

Central research question: Can expert knowledge (here: elements of bilateral symmetry of the lung fields in CXR scans) be leveraged to mitigate current challenges in medical image analysis, and if so, how?

Segmentation and Analysis of Lumbar Spine MRI Scans for Vertebral Body Measurement

This chapter explores the integration of expert knowledge as contextual information, focusing on lumbar spine MRI scan analysis. Clinicians often rely on specific measurements, such as vertebral body height, to identify spinal anomalies. However, manual analysis can be time-consuming and prone to inconsistencies [123, 124]. To address this, we propose a context-aware solution for lumbar spine MRI analysis that automates distance measurements. The proposed context-aware method comprises a DL-based segmentation of the relevant lumbar vertebrae and a rule-based expert knowledge system that utilizes the developed segmentation mask to measure pertinent distances. This approach results in a highly interpretable solution, as all final calculations are based on the visualized segmentation mask. Any incorrect distance measurements can be easily identified, allowing for manual correction by an expert.

We demonstrate that expert knowledge is not only beneficial for analyzing CXR data but is also highly applicable to other use cases within the medical imaging domain, regardless of the image modality. Furthermore, it emphasizes that *contextual information can be integrated in a two-stage process*, where the output of a data-driven segmentation model serves as the foundation for an expert knowledge system. This contrasts with the two previous Chapters 2 and 3, where expert knowledge is exclusively integrated into an end-to-end training process.
Please note that all experiments are based on a *German in-house data set* provided by the Evidia GmbH. Recent publications have explored DL-based analysis of lumbar spine MRI scans [130, 131, 132]. However, these studies have not considered German patient cohorts. Our research addresses this gap by evaluating the proposed context-aware method using German medical image data, offering unique insights into its performance within this specific healthcare system. This approach not only contributes to the broader field of medical image analysis but also provides locally relevant findings that could inform clinical practice in Germany. Detailed results and their implications are presented in Chapter 4, demonstrating the method's efficiency and potential impact on patient care.

Central research question: Can expert knowledge (here: distance measurement methods) be leveraged to mitigate current challenges in medical image analysis, and if so how?

Is One Label All You Need? Single Positive Multi-label Training in Medical Image Analysis

In contrast to Chapters 2, 3 and 4, which focus on expert knowledge as contextual information, the subsequent chapters concentrate on the contextual knowledge concerning the data set and label quality as essential insight. Specifically, this chapter addresses the challenge of missing labels in multi-label classification, with a particular focus on SPML training, one of the most challenging scenarios. We propose context-aware methods that enable high-performance DL training with only one observed positive label per sample, while all other labels remain unobserved. This research is particularly relevant for medical image analysis, where efficient SPML training can significantly reduce the costly and time-consuming annotation process for medical data sets. By leveraging the information about which labels are observed as contextual insight, our methods aim to improve classification accuracy and efficiency. We demonstrate the effectiveness of this approach using multi-label anomaly detection in CXR data as a practical use case.

To the best of our knowledge, this research represents the *first investigation of SPML training for the important medical image analysis domain*, thereby bridging the gap between state-of-the-art DL research and critical application areas. SPML training can attain classification results that are competitive with fully supervised training while exploiting significantly fewer labels, when focusing on computer vision benchmark data sets [63, 133]. However, we underline in our research that these results are not reproducible for the more complex use case of medical image analysis. Inspired by the remarkable results of the context-aware loss in Chapter 3, we therefore propose *novel context-aware loss functions that outperform state-of-the-art SPML methods* for the multi-label detection of anomalies in medical images. We integrate the contextual information, which labels are observed into the DL training, resulting in efficient implicit weighting factors during the gradient calculations for missing labels, ultimately enhancing the robustness of the DL method. Further results and a detailed introduction of the novel loss functions are outlined in Chapter 5.

Central research question: Is it possible to utilize contextual information (here: which labels are observed) to attain good performance in the multi-label classification of medical image data despite a substantial proportion of missing labels, and if so, how?

Improving Intensive Care Chest X-ray Classification by Transfer Learning and Automatic Label Generation

In this chapter, we revisit the concept of label quality as contextual information. In contrast to the previous SPML chapter, where each sample had missing labels, this chapter focuses on subsets of samples that exhibit varying label qualities. In addition to manually annotated scans (”gold labels”), we also examine samples with automatically generated annotations (”silver labels”), which usually lead to lower label quality. Automatic labels are generated based on the report corresponding to the image data (available for clinical routine data), leveraging a rule-based labeler [134]. For the given data set noisy labels remain unidentified. The contextual information regarding the label quality of these subsets represents a valuable prior insight for DL methods.

We aim to achieve high performance with reduced annotation effort. Two different strategies that incorporate contextual information into the modeling process are investigated. Firstly, we implement a transfer learning method based on public medical imaging data. Secondly, we propose a *context-aware pre-training strategy* that facilitates the efficient use of a silver-labeled subset. Following pre-training on the extensive silver labeled subset, the model is fine-tuned on the gold-annotated samples. Our findings suggest that context-aware pre-training outperforms transfer learning on public data when limited gold training data ($N \leq 5,000$) is available.
This research represents the *first utilization of an in-house multi-label ICU CXR data set provided by the University Hospital Bonn* to analyze DL-based DDSS. It therefore contributes to the investigation of context-aware DL methods on real-life clinical routine data, aiming to bridge the gap between state-of-the-art DL and the highly relevant field of medical image analysis. Given the time-consuming and costly nature of annotating medical image data, exploring context-aware methods that enable the efficient use of automatically generated labels emerges as a critical research area. This field is particularly relevant for companies, clinics, and researchers seeking to develop data-driven DDSS based on clinical routine data. Further information about the given research is outlined in Chapter 6.

Central research question: Is it feasible to integrate contextual information about the label quality (here: subsets with different annotation quality) to enhance performance when handling data sets with label noise, and if so, how?

Development of Image-based Decision Support Systems utilizing Information Extracted from Radiological free-text Report Databases with Text-based Transformers

Motivated by the successful use of automatically generated labels extracted from free-text reports of clinical routine data in the previous chapter, we further investigate the potentials and limitations of extensive automatically labeled data sets for the development of DL-based DDSS. We use the clinical in-house ICU CXR data provided by the University Hospital Bonn. The contextual information aligns with the previous chapter, presenting two subsets of data with differing label qualities: manually and automatically generated labels. Please note that in this chapter, compared to Chapter 6, the automatically generated labels demonstrate superior label quality due to the utilization of a complex transformer-based NLP labeler [134]. Although the annotation method for each sample is known, it remains uncertain which labels are accurate.

Silver-labeled data sets are frequently employed to train DL models with minimal annotation effort [6, 135, 136]. Although gold samples are often generated for performance evaluations or data-driven training of automatic labelers, the benefits of considering additional gold annotated samples are not sufficiently researched [6, 135, 136]. This thesis introduces a context-aware pre-training strategy that enhances the effective use of automatically generated silver labels while incorporating a small subset of high-quality gold annotations. As in the previous chapter, silver annotations are used to pre-train the model, while gold samples are used for fine-tuning.
Our research results suggest that combining gold and silver samples, specifically through *context-aware pre-training* based on prior insights regarding the label quality, improves performance while handling automatically generated labels. It serves as a proof of concept that automatically generated labels (with high label quality) are suitable for the development of context-aware DL-based DDSS with significantly reduced annotation effort. Despite the remarkable performance increase obtained with

automatically generated labels, Chapter 7 additionally underlines these annotations include label noise. This emphasizes noise-robust DL methods as a direction for future work. We further outline relevant results and contributions in Chapter 7.

Central research question: Is it feasible to integrate contextual information about the label quality (here: subsets with different annotation quality) to enhance performance when handling data sets with label noise, and if so, how?

Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems

Building on the insight that noise-robust training can be crucial for processing data sets with automatically generated labels, we introduce a novel context-aware method to enhance the robustness of DL models to label noise. As in Chapters 6 and 7, we leverage contextual information about label quality in the data sets. While we acknowledge the presence of label noise, the specific mislabeled instances remain unknown. Drawing from our previous work, we assume access to an estimate of the expected noise ratio. Since an assessment of the performance of automatic/human annotators on an independent test set with high label quality is crucial for managing the annotation process, this information is often provided. This approach aims to improve model performance and reliability when training on data sets with potentially inaccurate labels, a common challenge in real-world applications.

A wide range of research proposed successful noise-robust DL methods, achieving remarkable performance despite label noise. However, most methods were introduced for general computer vision and ML problems and are often not sufficiently evaluated for real-life challenges [43, 44, 45, 46, 137, 138, 139]. Additionally, they focusing on high noise levels of up to 80% [44, 45, 46, 137, 138, 139], neglecting lower noise levels, despite their considerable practical relevance. We emphasize in our work, that these exceptional achievements are not generally reproducible for real-life complex medical image use cases with realistic lower noise levels.
This chapter therefore proposes a *novel context-aware loss function* to use medical image data sets with noisy labels for the development of an efficient DDSS. We aim to further bridge the gap between state-of-the-art DL and the highly relevant application field of medical image analysis. The introduced loss function enables the model to abstain potential noisy samples during training. The estimation of expected label noise represents a valuable contextual information to guide the abstaining process. While enhancing noise robustness, the proposed loss function represents an easy-to-implement user-friendly DL method. In addition to noise simulation experiments on public data, enhanced noise robustness is demonstrated for a clinical in-house data set of ICU CXR scans with automatically generated labels provided by the University Hospital Bonn. These results underline the potential clinical impact of the proposed context-aware loss function, facilitating efficient noise-robust training for routine clinical data (with automatically generated labels). In addition to Chapter 3 and 5, this work highlights the ability of loss functions to incorporate contextual information into the data-driven training process.

Central research question: Is it feasible to integrate contextual information about the label quality (here: estimation of expected label noise ratio) to enhance performance when handling data sets with label noise, and if so, how?

CHAPTER 2

Towards Symmetry-aware Pneumonia Detection on Chest X-rays

The research discussed in the following chapter has been published in

H. Schneider, M. Lübbering, R. Kador, M. Broß, P. Priya, D. Biesner, B. Wulff, T. Bell Felix de Oliveira, Y.C. Layer, U. I. Attenberger, R. Sifa. **Towards Symmetry-aware Pneumonia Detection on Chest X-rays**. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022. http://doi.org/10.1109/SSCI51031.2022.10022222

2.1 Result Summary

The following chapter discusses the potential of expert knowledge as contextual information for the analysis of CXR scans.
A common machine learning approach to detect lung disease in CXR scans is to use deep CNNs and train them in a data-driven, supervised manner. While these solutions have demonstrated impressive performance across various applications [6, 126, 140], challenges such as model interpretability and data-efficient training persist. These issues can impede the progress of DDSS for new applications and their integration into clinical workflows, marking them as crucial areas for research in medical image analysis. In the subsequent chapters, we therefore introduce innovative context-aware DL methods designed to address these challenges by incorporating contextual information relevant to the analysis of CXR scans.
For instance, experts like radiologists use their prior knowledge to analyze this image data efficiently and accurately, often comparing the left and right lung fields intuitively [118, 141]. Although there is no perfect symmetry, for example, due to the shadow of the heart, elements of bilateral symmetry can be identified. The symmetrical elements are disturbed by various diseases and anomalies, experts therefore leverage pathological asymmetries during their diagnosis decision [118, 141]. While the elements of bilateral symmetry have demonstrated their value as expert knowledge for DL methods across various medical applications, their potential for CXR analysis has yet to be thoroughly investigated [127, 128, 129].
The following chapter aims to exploit this bilateral symmetrical information during the training of DL methods to enhance performance and enable data efficient training for the analysis of CXR data.

Specifically, we explore the analysis of CXR images of children younger than 5 years to automatically detect pneumonia, a significant factor in child mortality.

The main contributions are:

- To the best of our knowledge, the first investigation of the potential of context-aware DL methods based on the elements of bilateral symmetry of the lung fields in CXR images
- Proposition of novel context-aware model architectures, referred to as Symmetry-aware Merged and Symmetry-aware Concatenated Siamese Network, for CXR scan analysis,
- Evaluation of the enhanced performance, data-efficiency and interpretability of the proposed context-aware method.

Context-aware Method

We propose a novel symmetry-aware deep Siamese network for analyzing CXR scans, aiming to extract critical features from the lung fields independently. The backbone of this network is the DenseNet-121 architecture, which is trained using the CE loss. In the initial layers of the Siamese network, the weights of the first two dense blocks are shared and adjusted to obtain the hidden representations of the left and right lung fields separately. For this purpose, a CXR scan is separated into right and left lung fields , with the left lung being flipped to aid alignment during matching. We refer to the right and left input scans as $\boldsymbol{S}_{\text{right}}$ and $\boldsymbol{S}_{\text{left}}$. Subsequently, the generated features, which may hold crucial information about the asymmetry of the lung fields, are fused and further processed. Two fusion methods referred to as Symmetry-aware Merged Siamese Network (**SAM**) and Symmetry-aware Concatenated Siamese Network (**SAC**) are explored. In the SAC approach, the features from the left and right lung fields are fused through concatenation, ensuring no information loss. Conversely, the SAM method calculates the difference between the left and right lung field representations, $\boldsymbol{F}_{\text{left}}$ and $\boldsymbol{F}_{\text{right}}$, with the merging layer defined as follows

$$\boldsymbol{F}_{\text{merged}} = \boldsymbol{F}_{\text{left}} - \boldsymbol{F}_{\text{right}} \tag{2.1}$$

In the SAM model, an element-wise comparison of the left and right representations is performed, potentially capturing significant information about pathological asymmetries. Meanwhile, the SAC model focuses on extracting the most critical asymmetric information through data-driven training, ensuring no information loss in the fusion method.
Both model architectures do not increase the model complexity significantly. For the SAM network no additional parameters are used, the subsequent dense blocks and transition modules remain unchanged. For the SAC model, the backbone architecture requires slight adjustments to integrate the stacked feature maps. The concatenation occurs within a transition module, prior to the 1×1 convolution layer. Only the number of input parameters for this convolutional layer is doubled, while the rest remains unchanged. Consequently, the SAC increases the volume of trainable parameters by less than 2%. The architecture still allows the utilization of pre-trained weights from ImageNet, if desired. Figure 2.1 visualizes the proposed SAC approach.

Key Results

In [12] we present experimental results integrating a purely data-driven baseline utilizing the backbone DenseNet-121 as model architecture to evaluate the performance of the proposed context-aware

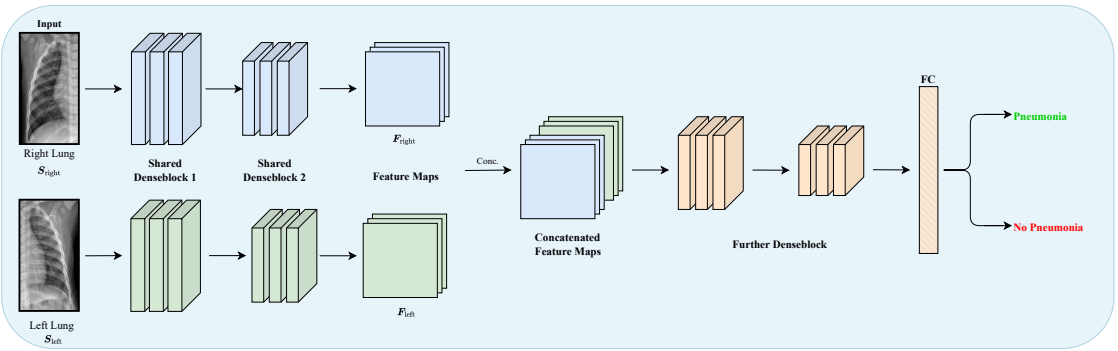


Figure 2.1: Representation of the context-aware architecture SAC. Right and left lung fields are initially processed separately by the Siamese network, which comprises the first two dense blocks of the DenseNet backbone. The obtained feature maps are concatenated and processed jointly by the downstream dense blocks and fully connected (FC) layers.

methods. Each architecture is initialized with five different seeds using Kaiming initialization [95] to provide confidence estimates on model performance. The reported scores represent the average of the five scores obtained for each model architecture. The enhancement of both accuracy performance and data efficiency through the integration of expert knowledge into network architecture is demonstrated. The proposed context-aware models achieve an improvement in the F1 score [142] by up to 2.0% without a significant rise in architectural complexity. When using only 25% of the training set, the SAC model experiences a mere 0.7% drop in the F1 score, compared to a 1.9% reduction in the purely data-driven training. This underscores the data efficiency of the proposed context-aware architectures, highlighting the advantage in leveraging expert knowledge effectively. The SAC model outperforms the SAM network, stressing the benefits of a data-driven fusion pipeline.

Furthermore, we investigate the interpretability of the context-aware methods by an in-depth analysis of the activation maps of context-aware and data-driven methods generated with the guided Grad-CAM algorithm. An increased interpretability of the methods is observed, as the context-aware algorithms focus more strongly on the regions of interest, the lung fields. The models also place an increased focus on non-pathological asymmetries, which are not pertinent to the diagnostic decision. Mitigating this effect by enhanced context-aware methods represents a future work. For a more comprehensive overview of the additional results, please refer to [12].

Overall, this chapter proposes context-aware DL methods enriched with expert knowledge regarding the elements of bilateral symmetry in lung fields of CXR scans. The analysis highlights the enhanced performance and data efficiency of the context-aware methods. Aspects that are particularly crucial for the development of DL-based DDSS, as the annotation of large-scale medical image data sets often presents considerable time and cost challenges. Additionally, an increased interpretability of the model is indicated. The proposed context-aware methods thereby contribute to the development of more efficient DDSS with enhanced clinical impact.

2.2 Author's Contributions

The pneumonia detection project was developed in cooperation with the University Hospital Bonn to generate the Pneumo.AI Demonstrator (https://pneumoai.ki.nrw/introduction). My contributions to the given research as first author are:
I was the main contributor of the research, planning and conceptualization of the context-aware CXR scan analysis method. Together with the second author M. Lübbering, I implemented the required DL training pipelines. All relevant experiments were performed by me. The paper was mainly written by me, with helpful regular meetings, discussions and proofreading from all co-authors.

CHAPTER 3

Symmetry-aware Siamese Network: Exploiting Pathological Asymmetry for Chest X-ray Analysis

The research summarized in this chapter has been published in the following paper

H. Schneider, E.C. Yildiz, D. Biesner, Y.C. Layer, B. Wulff, S. Nowak, M. Theis, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Symmetry-aware Siamese Network: Exploiting Pathological Asymmetry for Chest X-ray Analysis**. In *International Conference on Artificial Neural Networks (ICANN)*, 2023. https://doi.org/10.1007/978-3-031-44216-2_14

The research received the **Springer&ENNS Best Paper Award of the ICANN conference 2023**.

3.1 Result Summary

After the initial positive research results regarding the prior knowledge of the elements of bilateral symmetry of lung fields in CXR images for context-aware DL methods, we further explore this contextual information to tackle current challenges in the medical image analysis field. We propose context-aware solutions that amplify the impact of contextual information within the DL pipeline in order to fully exploit the potential of the given expert knowledge. We focus on enhancing the interpretability of the employed model while achieving exceptional performance through context-aware techniques. As outlined in the preceding chapter, these research areas represent significant challenges in the domain of medical image analysis, given their critical relevance to the integration of DL-based DDSS within the clinical workflows.
While in Chapter 2 the symmetry information is included through the architecture, within this chapter, an additional symmetry-aware loss function is proposed. We aim to enhance the impact of prior information based on expert knowledge on the training process. For this approach, in addition to conventional classification annotations, we utilize information about the location of the pathological asymmetries of each lung field. These areas are represented by bounding polynomials, visualized in Figure 3.2, which we refer to as the ground truth mask $\boldsymbol{Y}$. Note that pathological asymmetric regions

are also referred to as diseased areas in the following work.
We focus on the clinically highly relevant binary classification of healthy vs. non-healthy individuals, forming the basis of a data-driven prioritization system. These DL-based support devices can structure the clinical workflow more efficiently, as patients with potential anomalies require more urgent medical care.

The main contributions of the research are:

- Introduction of context-aware architectures, named Symmetry-Aware Siamese Networks, for the analysis of CXR images,
- Proposition of context-aware contrastive loss, integrating contextual information about pathological elements of bilateral asymmetry of the lung fields in CXR scans,
- To the best of our knowledge, the first investigation of symmetry-aware loss functions for CXR scan analysis,
- Evaluation of enhanced performance and interpretability of proposed context-aware models.

Context-aware Method
Inspired by the preceding Chapter 2 and [143], we introduce the Symmetry-Aware Siamese Network (**SASN**) as an end-to-end DL model with a DenseNet-121 backbone. It comprises three distinct modules: siamese encoding, feature fusion and feature comparison.

The **siamese encoding module** encodes the scan input $\boldsymbol{S}$ and its flipped form $\boldsymbol{S}_{\mathrm{f}}$, utilizing a siamese structure with two streams and shared weights. The calculated embeddings $\boldsymbol{F}$ and $\boldsymbol{F}_{\mathrm{f}}$ are further processed by the two proceeding modules.
The **feature fusion module** concatenates the generated feature maps in order to further process the information in a data-driven manner, building on the SAC network discussed in Chapter 2. This module produces a probability output mask $\hat{\boldsymbol{Y}}$ which is used for the final anomaly classification. During training, we apply an element-wise BCE loss between the probability output mask $\hat{\boldsymbol{Y}}$ and the ground truth $\boldsymbol{Y}$, representing diseased areas of the lung fields. This loss is defined as

$$L_{\mathrm{b}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = -\frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{Y}_i \log(\hat{\boldsymbol{Y}}_i) + (1-\boldsymbol{Y}_i)\log(1-\hat{\boldsymbol{Y}}_i)\Big), \tag{3.1}$$

for a ground truth mask with n pixels. A scan is classified as diseased during inference, if any region of the probability map contains a pixel value larger than a defined classification threshold.
The **feature comparison module** targets pathological asymmetries within the image data and facilitates the application of a symmetry-aware contrastive loss function. An element-wise L2 distance is computed between the generated feature maps after applying a non-linear transformation g. The objective is to minimize the element-wise distance between healthy (symmetric) areas while maximizing distance between pathological asymmetric regions. The context-aware loss is defined as

$$L_{\mathrm{c}}(\boldsymbol{F}, \boldsymbol{F}_{\mathrm{f}}) = \begin{cases} \sum_x \left\| g\Big(\boldsymbol{F}(x)\Big) - g\Big(\boldsymbol{F}_{\mathrm{f}}(x)\Big) \right\|^2 & \text{if } x \notin \hat{\boldsymbol{M}} \\ \sum_x \max\left(0,\ m - \left\| g\Big(\boldsymbol{F}(x)\Big) - g\Big(\boldsymbol{F}_{\mathrm{f}}(x)\Big) \right\|^2\right) & \text{if } x \in \hat{\boldsymbol{M}}, \end{cases} \tag{3.2}$$

where x represents a single pixel coordinate, while $\boldsymbol{F}$ and $\boldsymbol{F}_{\mathrm{f}}$ signify the embedded features of $\boldsymbol{S}$ and $\boldsymbol{S}_{\mathrm{f}}$, respectively. Areas marked as diseased in the target image mask are denoted by $\hat{\boldsymbol{M}}$ and we define the margin m as the radius for the dissimilarity of pathological asymmetries. This context-aware loss function is designed to enable the model to distinguish between pathological and non-pathological asymmetries. The generated distance enables the observation of the model's symmetry- (and context-) awareness. The overall established loss function is a combination of the context-aware and classification loss

$$\hat{L} = L_{\mathrm{b}} + L_{\mathrm{c}}. \tag{3.3}$$

Motivated by Chapter 2, we investigate the influence of a separate processing of the lung fields ($\mathrm{SASN}_{\mathrm{split}}$) compared to processing the whole image ($\mathrm{SASN}_{\mathrm{vanilla}}$). The $\mathrm{SASN}_{\mathrm{vanilla}}$ model is visualized in Figure 3.1. While both models depend on the general structure of the introduced siamese encoding, feature fusion, and feature comparison modules, further information about the adapted network architecture based on the input data is available in [13].

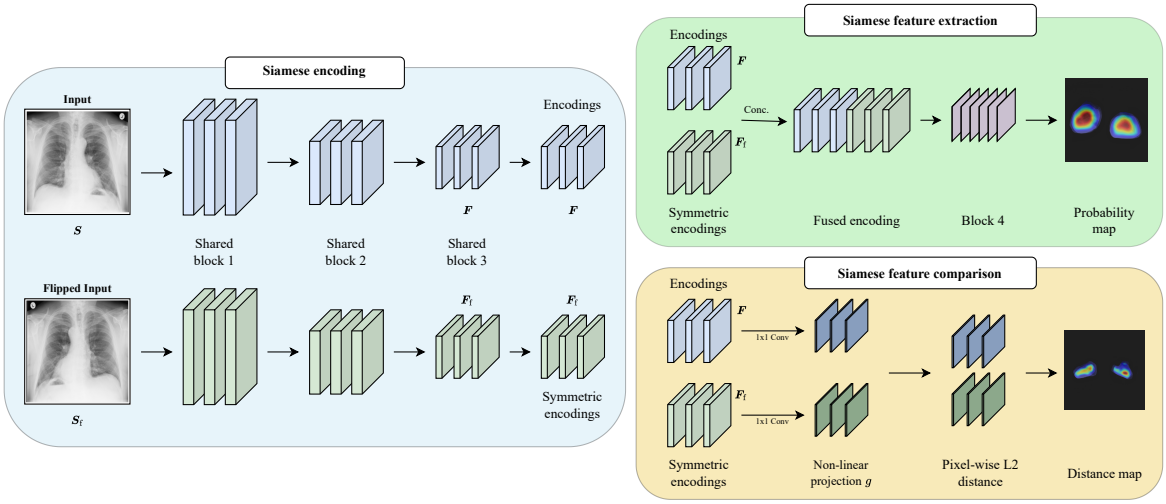


Figure 3.1: Model architecture of the $\mathrm{SASN}_{\mathrm{vanilla}}$. A CXR image and its flipped form are processed by an encoding module consisting of dense blocks with shared weights. The resulting embeddings are decoded by a siamese feature fusion module and a siamese feature comparison module to provide a disease probability map and distance map of the features, respectively.

Key Results

In [13] we present experimental results considering data-driven baselines, CheXNet [11] and vanilla Mask Region-Based Convolutional Neural Network (R-CNN) [144]. The performance is evaluated leveraging the measurements Compute Area Under the Receiver Operating Characteristic Curve (AUROC), F1 score and Average Precision [142]. We utilize 95% bootstrap confidence intervals based on 100 resamples. Non-overlapping confidence intervals indicate a significant improvement. We show that the context-aware methods can significantly improve the classification performance compared to data-driven baselines by up to 9.9% regarding the AUROC measurement. Given the potential impact of DL-based DDSS on patient outcomes, we strive to attain the highest possible

performance. The enhanced performance, rooted in expert knowledge, underscores the significance of the proposed context-aware DL method for medical image analysis.

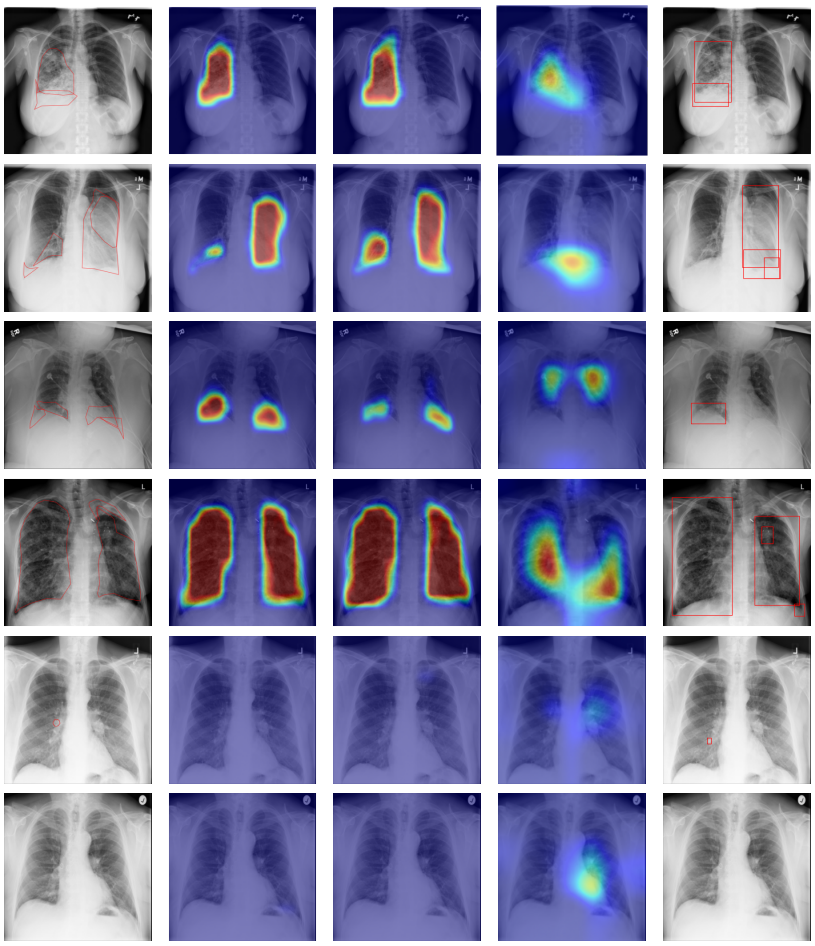
Additionally, we assess the interpretability of the context-aware modell by in-depth analysis of the visual output. The results indicate an improved interpretability of the proposed context-aware solution compared to other well-established interpretable methods for the baselines, visualized in Figure 3.2. The context-aware models effectively focus on diseased (and asymmtric) areas for their final classification. We demonstrate that the enhanced SASN models successfully target pathological asymmetries while disregarding non-pathological ones. This highlights the benefits of the proposed $\text{SASN}_{\text{vanilla}}$ and $\text{SASN}_{\text{split}}$ networks, which are trained with a context-aware loss function, in contrast to the SAM and SAC models previously discussed in Chapter 2.

The given chapter emphasizes that the inclusion of contextual information enhances both interpretability and performance, which are crucial attributes for DL methods in medical image analysis. Specifically, we underline that expert knowledge regarding the elements of the bilateral symmetry of the lung fields within CXR scans can be integrated into the model architecture and the loss function, thereby increasing the influence of contextual information on the training pipeline.

To summarize, Chapter 2 and 3 demonstrate the potential of expert knowledge as contextual information in medical image analysis. By considering the elements of bilateral symmetry of the lung fields in CXR scans during the DL training process, not only the performance but also the data efficiency and interpretability of the methods is improved. These issues are particularly relevant for the medical image analysis field.

3.2 Author's Contributions

My contributions to this paper as first author (shared first authorship) are as follows:
I was the main contributor of the research, planning and conceptualization of the context-aware CXR scan analysis method. I was responsible for the evaluation of the efficiency of developed algorithms and the generation of figures and tables in the final paper in equal parts with my co-author. I was the main contributor responsible for the writing of the paper, with helpful support through regular meetings, discussions and proofreading from all co-authors.



(a) (b) (c) (d) (e)

Figure 3.2: Visual outputs for different models. (a) CXR image with ground truth diseases, (b) probability map of $\text{SASN}_{\text{vanilla}}$, (c) probability map of $\text{SASN}_{\text{split}}$, (d) heat maps for CheXNet [11] generated with GradCam++, (e) Mask R-CNN [144] object detection boxes. For the generated masks, regions likely to contain diseased features are highlighted in red, whereas purple pixels signify a low probability. Both proposed methods $\text{SASN}_{\text{vanilla}}$ and $\text{SASN}_{\text{split}}$ focus solely on pathological asymmetries, highlighting their advantages over the SAC and SAM models discussed in the preceding Chapter 2.

CHAPTER 4

Segmentation and Analysis of Lumbar Spine MRI Scans for Vertebral Body Measurements

The research discussed in the following chapter has been published in

H. Schneider, D. Biesner, A. Ashokan, M. Broß, R. Kador, S. Halscheidt, B. Bagyó, P. Dankerl, H. Ragab, J. Yamamura, C. Labisch, R. Sifa. **Segmentation and Analysis of Lumbar Spine MRI Scans for Vertebral Body Measurements**. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2023. https://doi.org/10.14428/esann/2023.ES2023-88

4.1 Result Summary

Chapters 2 and 3 underline that expert knowledge about the use case can provide valuable information for data-driven solutions. However, contextual information is solely considered during the training process of the DL methods. In this chapter, we explore how expert knowledge can be integrated into our DL method through a two-stage process. We propose a rule-based expert knowledge system, that is applied on top of a DL-based segmentation method for the final decision support. This research highlights the various ways contextual information can be incorporated into DL-based solutions. Unlike the previous chapters, we do not focus on CXR data. Instead, we concentrate on the analysis of lumbar spine MRI scans, supplemented with additional contextual insights. This underscores the significance of context-aware DL methods across different use cases, irrespective of the image modality involved.
The interpretation of lumbar spine MRI scans often includes manual distance measurements, a time-consuming task requiring experienced radiologists and orthopedists. The growing volume of scans to be analyzed therefore represents a challenge for the everyday workflow of clinics and private practices. In addition, there is a high level of intra- and interreader variability, even among specialists, which can have a negative impact on the quality of reporting [123]. Thus, the automatic measurement of critical distances, such as the height of individual vertebral bodies has the potential to assist physicians in detecting anomalies and optimizing processes.

The main contributions are:

- Introduction of context-aware solution approach to automatically analyze lumbar spine MRI scans based on expert knowledge,
- Investigation of the potential of DL-based models for the segmentation of spinal vertebrae leveraging in-house volumetric MRI data from the German radiology company Evidia GmbH.

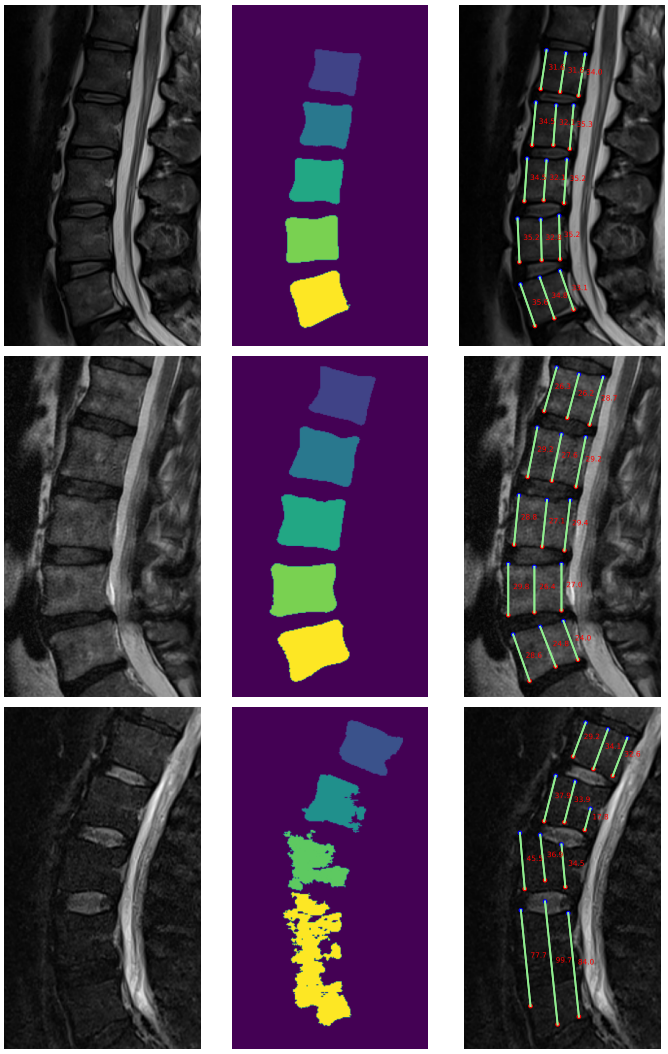
Context-aware Method
The aim of the following work is to measure the most important distances of lumbar spine MRI scans automatically. We develop an expert knowledge system, utilizing segmentation masks of the relevant vertebral bodies, generated by a DL model. We implement the SpineParseNet architecture [132] for the data-driven segmentation. The architecture is based on a combination of a 3D and 2D segmentation model. The down-scaled 3D scan is segmented through a CNN architecture. The high-resolution MRI scan is processed per slice through a 2D Residual U-Net architecture to calculate a final high-resolution segmentation of the vertebrae [132]. The low-resolution segmentation slice of the 3D model is considered as additional input. We investigate two different segmentation approaches. The binary classification differentiates between vertebral bodies and background, while the multi-class classification methodology facilitates the prediction of the vertebral bodies through the application of nine distinct vertebrae labels (thoracic T11-T12, lumbar L1-L5 and sacral S1-S2).

Based on expert knowledge, rule-based computer vision modules are created to automatically measure the relevant distances of vertebral bodies. In collaboration with experts at the Evidia GmbH, six pertinent anchor points per vertebra are defined. No additional data-driven training is necessary for the final distance measurements, as all anchor points can be established in a rule-based manner. The calculations use the smallest bounding trapezium of the automatically generated segmentation mask of each vertebra. All experiments and evaluations are conducted using a German in-house data set provided by Evidia GmbH, which comprises 142 volumetric T2-weighted sagittal lumbar spine MRI scans, annotated by two medical research assistants under the guidance of a radiology resident.

Key Results
In [14] we demonstrate the high segmentation performance for the pertinent vertebral bodies. The binary segmentation model achieves only slightly higher performance compared to the mutli-class segmentation approach, considering the Dice Score and Jaccard Index as measurements [145]. Since the associated class represents valuable information, the following analysis is based on the multi-class segmentation model. Figure 4.1 visualizes the pertinent steps of the introduced context-aware methods. The DL model achieves a highly accurate segmentation of the relevant vertebrae, attaining a Dice score of up to 96.6%. The proposed rule-based distance measurement system, which relies on the automatically generated masks of the vertebrae, produces well-calculated distances. It is important to note that this context-aware solution results in a highly interpretable algorithm, as all final calculations are grounded in the visualized segmentation mask. Any incorrect distance measurements can be directly identified, allowing for manual correction by an expert.

In [14], we demonstrate that expert knowledge is not only advantageous for the analysis of CXR data but also highly relevant for other use cases within the medical imaging domain, regardless of the image modality employed. This chapter is in contrast to the two Chapters 2 and 3, in which the expert knowledge was solely integrated into the data-driven training process. It underlines that contextual information can be integrated in a two-stage process, the output of a data-driven segmentation model

represents the foundation for an expert knowledge system.
Additionally, the experiments conducted in this research are based on a *German in-house data set* provided by the radiology company Evidia GmbH. While current publications explore DL-based analysis of lumbar spine MRI scans, it is important to note that the used data does not represent a German patient cohort [130, 131, 132]. This research therefore offers a highly pertinent evaluation of the proposed context-aware method using German medical image data.



(a) (b) (c)

Figure 4.1: Visualization fo the most important steps of our context-aware solution. (a) Original scan, (b) predicted segmentation, (c) automatic measurements of vertebrae height. The segmentation of the first two scans is precise, which leads to highly accurate distance measurements. All distances are calculated in millimeter. The third scan shows less contrast and is more difficult to segment. This results in inaccurate distance measurements. Manual correction of the calculated distances is required.

4.2 Author's Contributions

My contributions to this paper as first author (shared first authorship) are as follows:
I collaborated on the research, planning, and conceptualization of the project equally with my co-author. I primarily oversaw the data acquisition, pseudonymization, annotation, and implementation of the data infrastructure. I contributed to the implementation of the context-aware methods and the tuning of the DL training equally with my co-author D. Biesner. I was responsible for the generation of the figures and tables in the final version of the paper. Together with my first co-author, I was equally responsible for the writing of the paper, with helpful support through regular meetings, discussions and proofreading from all co-authors.

CHAPTER 5

Is One Label All You Need? Single Positive Multi-label Training in Medical Image Analysis

The research discussed in the following chapter has been published in

H. Schneider, P. Priya, D. Biesner, R. Kador, Y.C. Layer, M. Theis, S. Nowak, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Is One Label All You Need? Single Positive Multi-label Training in Medical Image Analysis**. In *IEEE International Conference on Big Data (BigData)*, 2023. https://doi.org/10.1109/BigData59044.2023.10386758

5.1 Result Summary

The previous chapters discuss expert knowledge as contextual information in medical image analysis. The forthcoming chapters concentrate on the prior knowledge concerning the data set and label quality as essential contextual insights. Consideration is given to aspects such as missing labels and label noise, which are particularly relevant for the analysis medical image data sets. This contextual information offers valuable insight for data-driven training, as these data set characteristics can adversely affect the performance of the network. By addressing these challenges, the study aims to enhance the performance and robustness of DL methods.

First, we analyze the impact of context-aware methods for SPML training, introduced in Section 1.2. The contextual insight into which observed labels are available for a sample in the data set presents valuable information during the loss calculations. We therefore propose context-aware loss functions, aiming to achieve exceptional performance with less annotation effort and missing labels. A contribution that is particularly valuable for the often expensive and time-consuming annotation of medical image data sets. We focus on the multi-label analysis of CXR scans as a clinical use case.

The main contribution are:

- To the best of our knowledge, the first evaluation of state-of-the-art SPML training for medical image analysis,
- Introduction of novel context-aware SPML loss function, called Implicit Weighting Assume

Negative and Generalized Assume Negative Loss, that outperform state-of-the-art SPML loss functions.

Context-aware Method

As discussed in Section 1.2, the fully observed label vector is unknown for SPML training samples. For each sample only one positive and no negative label is observed. One solution how to handle the unobserved labels is to assume they are negative and to train the model with the conventional BCE loss, since in a typical multi-label data set negative labels are far more represented than positive labels. We refer to this approach as the assume negative (AN) method. However, the introduced false negative labels can harm the performance, since the BCE loss is not inherently noise-robust, outlined in Section 1.2. We therefore propose novel context-aware loss functions, Implicit Weighting Assume Negative Loss (**IWAN**) and Generative Assume Negative Loss (**G-AN**), to reduce the influence of introduced false negative labels on the training process.

The objective of the IWAN loss is to incorporate an implicit weighting factor during gradient calculations that mitigates the impact of potentially noisy labels, while maintaining high performance. The loss function is based on the assumption that a high model output associated with an unobserved, presumed negative label is likely indicating a false negative label. Consequently, it is defined as follows

$$\begin{aligned} L_{\text{IWAN}}\Big(f(\boldsymbol{x}), \boldsymbol{z}\Big) &= -\frac{1}{2}\sum_{i=1}^{C}\Big[\boldsymbol{z}_i \log\Big(f_i(\boldsymbol{x})\Big) - (1-\boldsymbol{z}_i)\gamma\frac{1}{p}f_i(\boldsymbol{x})^p\Big] \\ \frac{\delta L_{\text{IWAN}}\Big(f(\boldsymbol{x}), \boldsymbol{z}\Big)}{\delta \boldsymbol{\theta}} &= -\frac{1}{2}\sum_{i=1}^{C}\Big[\boldsymbol{z}_i\frac{1}{f_i(\boldsymbol{x})}\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x}) - (1-\boldsymbol{z}_i)\gamma\frac{1}{f_i(\boldsymbol{x})^{(1-p)}}\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x})\Big], \end{aligned} \tag{5.1}$$

where $\boldsymbol{z}$ represents the label vector, for which unobserved labels are mapped to 0, $\boldsymbol{x}$ is the corresponding input sample, C is the number of considered classes and p and γ represent additional weighting parameters. The IWAN loss results in an updated implicit weighting factor $f_i(\boldsymbol{x})^{-(1-p)}$ during the gradient calculations for unobserved labels. If the model output is close to the considered negative label, we assume that it is a correct annotation. The implicit weighting factor increases, leading to a higher influence of the potentially correct sample.
The parameter $p \in (0, 1]$ rules how strongly we want to focus on the assumed clean labels during training. Lower p values result in a strong influence of the assumed clean labels. However, p values that are to low may hinder the learning process, as challenging true negative examples only have a limited influence on the training dynamics. In SPML training, setting $\gamma = 1/(1-C)$ leads to the single observed positive label having the same weighting influence on the loss function as the $C-1$ assumed negative labels. Further details on the IWAN loss function and the introduced parameters are discussed in [15].

Additionally, the G-AN loss function is designed to mitigate the impact of noisy labels while preserving high performance levels. Inspired by the generalized cross-entropy loss function, the G-AN is defined as follows

$$
\begin{aligned}
L_{\text{G-AN}}\Big(f(\boldsymbol{x}), \boldsymbol{z}\Big) &= -\frac{1}{2}\sum_{i=1}^{C}\left[\boldsymbol{z}_i \log\Big(f_i(\boldsymbol{x})\Big) - (1-\boldsymbol{z}_i)\gamma\frac{1-\Big(1-f_i(\boldsymbol{x})\Big)^q}{q}\right] \\
\frac{\delta L_{\text{G-AN}}\Big(f(\boldsymbol{x}), \boldsymbol{z}\Big)}{\delta \boldsymbol{\theta}} &= -\frac{1}{2}\sum_{i=1}^{C}\left[\boldsymbol{z}_i \frac{1}{f_i(\boldsymbol{x})}\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x}) - (1-\boldsymbol{z}_i)\gamma\frac{1}{\Big(1-f_i(\boldsymbol{x})\Big)^{(1-q)}}\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{x})\right],
\end{aligned}
\tag{5.2}
$$

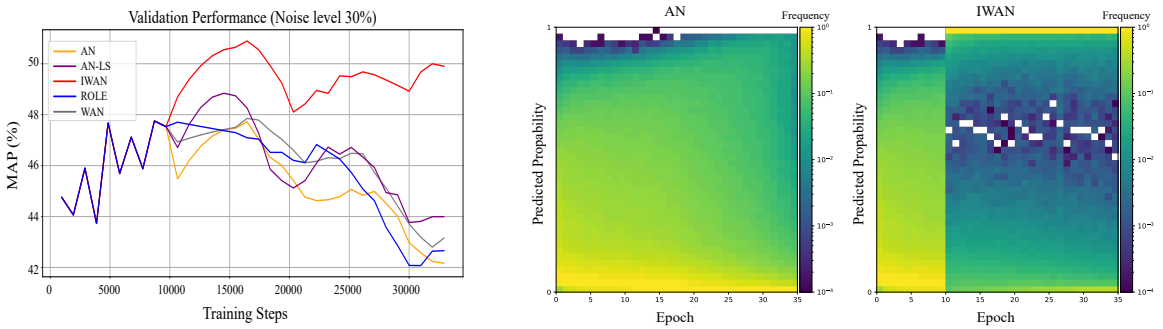
with $q \in (0, 1]$. The gradients for unobserved label are weighted by the factor $(1 - f_i(\boldsymbol{x}))^{-(1-q)}$. In contrast to the BCE loss, this approach reduces the emphasis on examples exhibiting low agreement between the prediction and the assumed target. The choice of parameter q represents a trade-off between noise-robustness and good learning dynamics. Higher q values result in an enhanced noise-robustness. However, q values that are too high may hinder the training process. To further diminish the impact of potentially noisy samples, we establish $\gamma = 1/(1 - C)$. A comprehensive introduction of the G-AN loss can be found in [15]. Note that the distinction between the IWAN and G-AN loss is rooted in their underlying motivations of the implicit weighting factors. In the case of IWAN, greater emphasis is placed on the most likely clean labels that are near the assumed target of 0, aiming to reduce the influence of noisy labels. Conversely, the G-AN loss allocates stronger weights to challenging samples whose predictions diverge from the assumed target of 0.

Key Results

In [15] we conduct experimental studies to investigate the potential of SPML loss functions for medical image analysis, leveraging the mean average precision (MAP) and AUROC [15, 146] as performance measures. Among others, we compute the 95% confidence intervals using bootstrapping with 1000 resamples to assess performance more accurately. Non-overlapping confidence intervals are interpreted as a significant improvement.

The proposed context-aware loss functions improve SPML training compared to several state-of-the-art baselines on different data sets for the analysis of CXR images. A performance increase of maximum 4.6% regarding the MAP can be achieved compared to the common AN solution. Even more complex SPML methods, such as the assume negative label smoothing baseline [63], can be outperformed by up to 2.5% in terms of MAP measurements. For an extensive training data set ($N = 123,801$) the IWAN loss surpasses the G-AN training. Figure 5.1 further underlines the remarkable performance of the introduced SPML loss functions. The given results highlight the potential of the proposed context-aware methods for SPML training in the medical image domain and represents a valuable baseline for further advancements of SPML training for medical image analysis.

State-of-the-art research highlights that SPML training can achieve performance comparable to fully supervised training, while requiring significantly fewer labels when focusing on computer vision benchmark data sets [63, 133]. Our initial investigation of SPML training for medical image data demonstrates that for a complex use case the highest performance is obtained leveraging fully observed labels, significantly outperforming all state-of-the-art SPML loss functions. This underlines the importance of investigating state-of-the-art DL methods for medical image data to enhance their evaluation and applicability to real-world challenges. In addition to proposing novel SPML loss functions, our research thus connects cutting-edge DL research with the critical domain of medical image analysis.



(a) (b)

Figure 5.1: Result analysis of the SPML training: (a) Validation curves during training with state-of-the-art SPML loss functions, introduced in [63]. The first 100 epochs involve a warm-up training with the AN loss. The proposed IWAN loss demonstrates the least tendency of overfitting. (b) Distribution of predicted probabilities for unobserved positives during SPML training. Each column represents a normalized histogram, white pixels emphasize a zero frequency. Training with IWAN (right) leads to the recovery of a significant number of unlabeled positives. The majority of the probability is correctly concentrated at 1 (top right) by the end of training. For AN (left) this behavior is not observed. Similar results are obtained for the G-AN training, outlined in [15].

The presented results emphasize that contextual information regarding label quality is invaluable for the processing of medical image data sets. In addition to the insight provided in Chapter 3, we underscore that loss functions are particularly well-suited for integrating contextual information during the training process. This assertion holds true not only for expert knowledge as a form of contextual information but also regarding prior insights concerning label quality.

5.2 Author's Contributions

As a first author I was responsible for the research, planning, and conceptualization of the project. The proposed context-aware loss functions were developed by me. I implemented the required DL pipelines and conducted all relevant experiments. I was responsible for the analysis of the results, including all figures and tables in the final paper. The paper was mainly written by me, with helpful support through regular meetings, discussions and proofreading from all co-authors.

CHAPTER 6

Improving Intensive Care Chest X-ray Classification by Transfer Learning and Automatic Label Generation

The research discussed in the following chapter has been published in

H. Schneider, D. Biesner, S. Nowak, Y.C. Layer, M. Theis, W. Block, B. Wulff, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Improving Intensive Care Chest X-Ray Classification by Transfer Learning and Automatic Label Generation**. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2022. https://doi.org/10.14428/esann/2022.ES2022-85

6.1 Result Summary

In this chapter, the concept of the corresponding label quality as contextual information is revisited. Unlike the previous SPML Chapter 5, where several labels of one sample had missing annotations, this chapter deals with subsets of samples that exhibit varying label qualities. It is established that samples from one subset are more prone to label noise due to the annotation process. Alongside manually annotated scans, we consider samples with automatically generated annotations, which typically result in lower label quality. However, noisy labels remain unidentified. The contextual information regarding the label quality of these subsets is incorporated into the DL solution approach. This chapter marks the first effort to categorize a multi-label ICU CXR data set provided by the University Hospital Bonn, leveraging automatically generated labels. Given the time-consuming and costly nature of annotating medical image data, exploring context-aware methods that enable DL training with minimal manual annotation effort emerges as a critical research area for companies, clinics, and researchers alike, who want to develop data-driven DDSS based on clinical routine data.

The main contributions are:

- Initial processing of ICU CXR data set provided by the University Hospital Bonn,
- Investigation of the context-aware methods, transfer learning based on public data and context-

aware pre-training based on a data subset with automatically generated labels, to handle limit manual annotations,

- Evaluation of the impact of different quantities of manually annotated data.

Context-aware Method

We define two groups of training samples for the clinical in-house data set: ”gold” and ”silver” annotated scans. Annotations are derived based on the corresponding text reports of the images, available for clinical routine data. Gold labels are manually extracted, while silver labels are obtained using a rule-based NLP approach. The rule-based labeler seeks out specific terms, negations and descriptions of uncertainty, while also applying additional text-based rules within the ”findings” section of the report to generate annotations [134]. Since the report texts describe the corresponding image data, we generate ground truth data not just for the report but also for the corresponding scan. Further details on the label extraction are discussed in [134]. Due to the error rate of the rule-based NLP labeler, silver samples are more likely to include noisy labels. However, this method enables the generation of extensive annotated data sets with minimal effort, facilitating training with large data sets at reasonable annotation costs.

To efficiently utilize the silver samples we propose a context-aware pre-training strategy. We first train the network on silver data, followed by a fine-tuning on the high-quality gold samples. This enables the inclusion of additional information based on the extensive silver data set, while mitigating the effects of noisy labels.

Moreover, a common context-aware method to include additional information in the training process is transfer learning based on publicly available data, which is introduced in Section 1.2. We pre-train the network to extract valuable information regarding the public data set, followed by a fine-tuning on gold samples. It is important to note that while both the public and our in-house data set consist of CXR images, they differ in their acquisition contexts. The public data images are obtained in a standardized imaging setting with patients in an upright position. In contrast, the in-house data set includes only anteroposterior thoracic images acquired in the ICU using a portable chest radiograph, with patients lying down in a hospital bed.

In addition, we combine both context-aware methods by first pre-training on a public data set, further training on the extensive silver data set, to finally fine-tune the model on gold samples. The aim is to integrate valuable knowledge of the public and silver data into out model.

Key Results

In [16] we conduct an experimental study to evaluate the introduced solution approach, leveraging the AUROC score [146] for evaluation. As a baseline, we additionally train the model solely using gold samples. To simulate the impact of reduced manual annotation, we limit the volume of gold training data at regular intervals, ranging from $N = 500$ to the full number of $N = 15,835$ images. These experiments are particularly relevant to the medical imaging domain, where the creation of extensive manually annotated data sets poses significant time and cost challenges.

When only a limited number of manually annotated gold labels are available $(N \leq 5,000)$, the context-aware inclusion of the automatically generated silver label enhances performance, surpassing transfer learning on a public CXR data set. For instance, with only 500 manually annotated scans, the AUROC score improves by up to 11.1% through inclusion of the the silver labeled data set compared to solely gold training, outperforming transfer learning by 9.3%. This underlines that for a limited gold data set, the information extracted from the silver data set proves to be more valuable than the insights

based on public CXR data, with a different acquisition context. The combination of both context-aware methods yields the highest performance, highlighting the extraction of valuable information based on the silver annotated subset and the public CXR data. However, as the number of gold annotations increases, the transfer learning based on a public data set outperforms the use of the silver labeled data. Our results suggest that the advantages of the silver pre-training depend on the quality of the silver labels when a sufficient gold data set is available. If the automatically generated labels are too inaccurate, transfer learning serves as a valuable alternative, avoiding noise overfitting during the pre-training phase.

The given chapter represents a highly pertinent evaluation of state-of-the-art DL methods for German clinical routine data. Thereby, we bridge the gap between DL research and the complex medical image analysis domain, laying the groundwork for more sophisticated and clinically impactful applications of these technologies in healthcare.
In addition, the presented work emphasizes that context-aware methods taking label quality into account are advantageous, not only for addressing missing annotations, but also for managing automatically generated labels. By incorporating contextual information about the annotation process into DL training, we facilitate the efficient utilization of automatically generated labels. In contrast to the previous chapters, we demonstrate that contextual information regarding label quality can be effectively integrated into the training pipeline through context-aware pre-training.

6.2 Author's Contributions

My contributions to this paper as first author (shared first authorship) are as follows:
I collaborated equally with my co-author on the planning and conceptualization of the project. We equally contributed to the implementation of the context-aware methods. I conducted all relevant experiments. Additionally, I was responsible for the generation of the result table in the final version of the paper and, together with my first co-author, was equally responsible for the writing of the paper, with helpful support through regular meetings, discussions and proofreading from all co-authors.

CHAPTER 7

Development of Image-based Decision Support Systems utilizing Information Extracted from Radiological Free-text Report Databases with Text-based Transformers

The research discussed in the following chapter has been published in

S. Nowak, H. Schneider, Y.C. Layer, M. Theis, D. Biesner, W. Block, B. Wulff, U.I. Attenberger, R. Sifa, A.M. Sprinkart. **Development of Image-based Decision Support Systems utilizing Information Extracted from Radiological Free-text Report Databases with Text-based Transformers**. In *European Radiology*, 2024. https://doi.org/10.1007/s00330-023-10373-0

Please note that this represents a shared first authorship.

7.1 Result Summary

Within this chapter, we further investigate the context-aware processing of the clinical in-house data set of ICU CXR data provided by the University Hospital Bonn. We explore the potential of automatically generated silver labels that exhibit higher label quality. To enhance this quality, we use a transformer-based NLP annotator. The contextual information aligns with the previous Chapter 6, presenting two subsets of data with differing label qualities: manually (”gold”) and automatically (”silver”) annotated scans. Although the annotation method for each sample is known, it remains uncertain which labels are accurate. Additionally, we further investigate the potential of context-aware pre-training to incorporate contextual information regarding label quality to improve performance while handling automatically generated labels. This area of research is critical for companies, clinics, and researchers seeking to develop data-driven DDSS, leveraging clinical routine data with minimal annotation effort.

The main contributions are:

- Exploration of the potential and the constraints of employing transformer-based report annota-

tions for the development of context-aware image-based DDSS,

- Strengthen analysis of context-aware pre-training strategy, introduced in Chapter 6,
- Examination of the impact of less extensive gold annotations on classification performance.

Context-aware Method

In this study, we utilize image data from the previous Chapter 6. The annotation primarily relies on the corresponding text report. We leverage manually ("gold") and automatically ("silver") generated labels. Since the report texts describe the corresponding image data, we generate ground truth data not just for the report but also for the corresponding scan. The silver labels are derived from transformer-based models trained on manually annotated gold labels for the reports, thus demonstrating higher label quality compared to the rule-based labels of the previous Chapter 6. We utilize Bidirectional Encoder Representations from Transformers (BERT) as an established transformer model to process the text reports. Please refer to [134] for more details on the applied NLP approach. In addition, all patients under the age of 16 are excluded from the following study. Only image data with a clear text-image mapping is considered. Furthermore, we generate an additional test set annotated based on the image data.

Silver-labeled data sets are utilized to train DL models with minimal annotation effort. While gold samples are frequently generated for performance evaluations or the data-driven training of automatic labelers, the advantages of incorporating additional gold-annotated samples remain underexplored [6, 135, 136]. In this thesis, we therefore investigate the potential of context-aware pre-training to efficiently leverage gold-annotated samples for the development of DL-based DDSS. We implement the following DL strategies:

- Training on the high-quality but limited gold samples set (M_{G}),
- Training on the extensive silver data (M_{S}),
- Pre-training on extensive silver data set followed by fine-tuning on high-quality gold samples $(M_{\mathrm{S}\setminus\mathrm{G}})$,
- One end-to-end training without considering the contextual information about the annotation process $(M_{\mathrm{S+G}})$. These runs are motivated by the high performance scores of the automatic labeler.

The experiments are conducted using simulations with varying volumes of gold-labeled data sets. It's important to note that the transformer models used for report content classification (generation of silver labels) also utilize varying amounts of manually gold-labeled reports, allowing for an evaluation of the overall impact of different levels of human annotation effort.

Figure 7.1 represents an overview of the entire study including the generation of the annotations, data sets and experiments.

Key Results

The following experimental results are based on the image-based test data set and the AUROC measurements for performance evaluation [146]. Significant differences were evaluated by non-overlapping 95% confidence interval calculated using bootstrapping with 1000 resamples.
Using transformer-based silver labels proves to be advantageous for developing image-based DDSS for

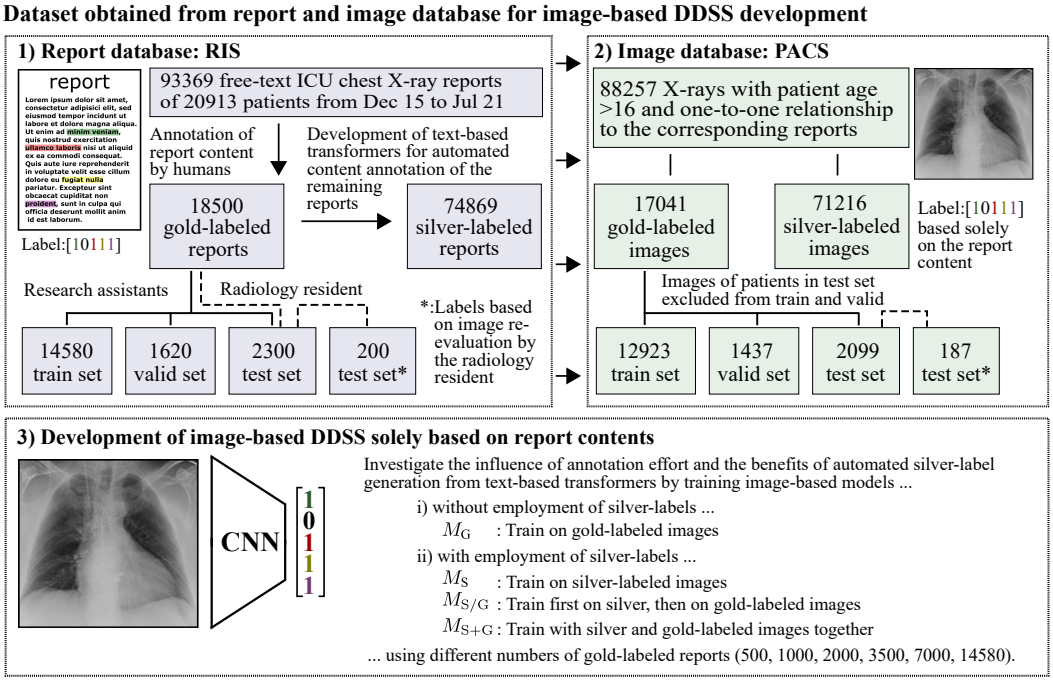


Figure 7.1: Summary of complete study. (i) We export the report contents of CXR examinations from ICU patients from the radiology information system (RIS). For a subset of the exported reports, the text content is annotated manually. The obtained gold labeled data set is divided into training, validation and test set. We reevaluate 200 samples of the test subset based on the image content to create image-based gold labels for testing and to investigate the disagreement to the report content. Text-based transformer models that automatically generate silver labels for the remaining reports are developed in a previous study using the gold-labeled reports. (ii) Images of patients older than 16 years with a clear one-to-one relationship to their associated reports are considered and extracted from the Picture Archiving and Communication System (PACS). Overall, we generate image data sets by utilizing the corresponding scans to different report data sets with either report content-based gold or silver labels or image-based gold labels. (iii) Based on these data sets we investigate different approaches for leveraging report content for the development of image-based DDSS.

ICU CXR examinations. All models leveraging the silver data set, M_S, M_{S+G} and $M_{S\setminus G}$, demonstrate significant improvement compared to solely gold data training M_G. A maximum performance increase of 13.3% is achieved by utilizing the generated silver data set. Additionally, the integration of gold-annotated samples enhances model performance compared to solely silver training by up to 2.5%, across all varying quantities of gold-annotated samples.
If more than 2000 manually annotated gold samples are available for training, the context-aware $M_{S\setminus G}$ model achieves the highest performance, surpassing the M_{S+G} training approach. This highlights the benefits of context-aware DL, which integrates contextual information about the annotation process and label quality during training. Furthermore, these results underline that even with the enhanced label quality of the silver annotations, the inherent noise in the silver data sets can negatively impacts model performance when implementing a purely data-driven training process. This emphasizes the necessity for noise-robust context-aware methods. Further result are outlined in [17].

In addition to Chapter 6, this work underlines that context-aware pre-training is a valuable tool

to integrate contextual information about the label quality in the DL training pipelines, even when leveraging a more powerful transformer-based annotator.
Furthermore, it demonstrates that automatically generated labels (with high label quality) are suitable for the development of context-aware DL-based DDSS. A research area crucial for companies, clinics, and researchers aiming to create data-driven DDSS based on clinical routine data with reduced annotation effort. Consequently, this work contributes to bridging the gap between state-of-the-art DL methods and the highly relevant medical image analysis domain, aiming to generate clinical impact with context-aware DL.

7.2 Author's Contributions

My contributions to this paper as first author (shared first authorship) are as follows:
I collaborated equally with my co-author on the research, planning and conceptualization of the project. I implemented the DL pipelines and was responsible for the fine-tuning of the proposed methods. We equally contributed to the generation of figures and tables in the final version of the paper and the writing of the text with helpful support through regular meetings, discussions and proofreading from all co-authors.

CHAPTER 8

Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems

The research discussed in the following chapter will be published in

H. Schneider, S. Nowak, A. Parikh, Y.C. Layer, M. Theis, W. Block, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems**. Accepted at *International Conference on Neural Information Processing (ICONIP)*, 2024.

A preprint version of the paper has been published in

H. Schneider, S. Nowak, A. Parikh, Y.C. Layer, M. Theis, W. Block, A.M. Sprinkart, U.I. Attenberger, R. Sifa. **Informed Deep Abstaining Classifier: Investigating Noise-robust Training for Diagnostic Decision Support Systems**. In *arXiv preprint arXiv:2410.21014*, 2024.

8.1 Result Summary

Building on the insights from the previous chapters, which highlight that utilizing automatically generated labels based on text reports may introduce label noise, we propose a context-aware, noise-robust training algorithm. As in the Chapters 6 and 7, the label quality serves as contextual information. It is known that the data set includes label noise, but the specific noisy labels remain unidentified. Furthermore, motivated by the prior chapters, it is assumed that an estimate of the expected noise ratio is accessible. This information is typically provided during the annotation process of clinical data, as an assessment of the performance of automatic/human annotators on an independent test set with high label quality is essential for managing the annotation process.
Unlike the previous Chapters 6 and 7, we do not consider data sets with varying annotation quality and therefore cannot rely on context-aware pre-training. However, the proposed method can be applied for efficient noise-robust (pre-)training using automatically generated labels. In the following sections, we further explore the potential of contextual information regarding label quality in

medical image analysis using DL. Specifically, this chapter presents an easy-to-implement, context-aware and noise-robust approach to utilize clinical routine data with automatically generated labels for the development of efficient DDSS. A field interesting not just for clinics but researchers and companies aiming to exploit the full potential of clinical routine data without extensive annotation effort.

The main contributions are:

- Proposition of the novel context-aware loss, referred to as Informed Deep Abstaining Classifier Loss, enabling the incorporation of prior knowledge on the expected noise ratio inside the DL loss function,
- Evaluation of noise-robustness of proposed and state-of-the-art loss functions on noise simulations, focusing on practically relevant lower noise level between 1% and 15% noise,
- Examination based on clinical routine data with automatically generated labels, highlighting the impact of the introduced context-aware method for the development of DDSS with automatically annotated image data.

Context-aware Method

As discussed in section 1.2, the CE loss is a commonly used classification loss that has achieved extraordinarily high performance for a wide range of multi-class applications, including medical imaging [8, 39]. However, it is not inherently noise-robust, as the performance can be significantly affected by label noise. Several solution approaches have been developed to maintain the high performance of the CE loss, while enhancing noise-robustness, such as the Deep Abstaining Classifier (DAC) [138]. The Informed Deep Abstaining Classifier (**IDAC**) loss introduced below is an extension of the DAC loss, integrating an estimation of the noise ratio $\tilde{\eta}$ of the given data set as additional contextual information.

The classifier used for training with the IDAC loss includes an additional $C+1$ output neuron that represents the probability of abstention for a sample. The parameter C indicates the number of true classes. This enables the model to abstain samples that are potentially noisy during training. The proposed context-aware IDAC loss is therefore defined as

$$\begin{aligned} L_{\text{IDAC}}\Big(f(\boldsymbol{x}), \boldsymbol{y}\Big) &= \Big(1 - f_{C+1}(\boldsymbol{x})\Big)\Big(- \sum_{i=1}^{C} \boldsymbol{y}_i \log\Big(\frac{f_i(\boldsymbol{x})}{1 - f_{C+1}(\boldsymbol{x})}\Big)\Big) + \alpha(\tilde{\eta} - \hat{\eta})^2 \\ \hat{\eta} &= \sum_{j=1}^{B} \frac{f_{j,C+1}(\boldsymbol{x})}{B} \end{aligned} \tag{8.1}$$

where $\boldsymbol{x}$ is the input sample, $\boldsymbol{y}$ is the corresponding label vector, B represents the batch size and $\hat{\eta}$ is the currently applied abstention of the classifier per batch. The parameter $\alpha > 0$ represents the abstention weight, a fixed hyperparemeter. The proposed IDAC loss consists of two terms. The first term represents an adaptation of the CE loss, which incorporates the abstaining probability prediction of the sample to level its influence during training. If the abstaining probability is equal to 0, this term is equivalent to the CE loss; if it is close to 1, the sample has little influence on the calculation. The aim is to learn the general features relevant for the given problem setup to enhance the classification

performance by reducing the impact of potential noisy samples.

The second term regularizes the abstention, preventing the classifier to abstain all samples. This regularization incorporates the contextual information of the label noise estimation, which can enhance the training process by providing valuable insight into how many samples should be abstained from during training. If the model abstains from too many samples, the regularization term increases, resulting in a stronger penalty for abstaining. This can lead to a decrease in the number of abstained samples, ideally excluding correct (but difficult to learn) training samples during the abstention process. Conversely, a low abstention rate compared to the expected noise estimation during training also results in a high regularization term. Minimizing this penalty term leads to an increase in the number of abstained samples. The aim is to exclude more noisy label samples during training, thereby reducing the effects of noise overfitting on the training performance. The regularization term is weighted by the abstention weight α, adjusting the influence of the abstention regularization during training.

Key Results

In [18] we conduct experiments, simulating different noise levels on publicly available CXR data, focusing on lower noise levels below 15% for the use cases pleural effusion and cardiomegaly. In addition, we implement the IDAC loss to process clinical routine data with automatically generated labels (a subset of the noisy silver data set from the previous Chapter 7). This facilitates an enhanced evaluation of the proposed context-aware loss function for real-life challenges. For the evaluation we consider the AUROC measurement [146]. We calculate the 95% confidence intervals using bootstrapping with $1,000$ resamples to enhance performance evaluation. Non-overlapping confidence intervals are interpreted as significant difference.

These experiments are driven by the observation that, while a broad spectrum of research has introduced successful noise-robust DL methods, these techniques were introduced for general computer vision and ML problems. Often, they are not sufficiently evaluated for real-life challenges, focusing on unrealistically high noise levels of up to 80% [43, 44, 45, 46, 137, 138]. Lower noise levels have not been sufficiently researched in the existing literature, despite their significant practical relevance. We demonstrate, that the remarkable noise-robust results for computer vision benchmark data sets are not generally reproducible for the more complex use case of medical image analysis. This highlights the necessity to evaluate state-of-the-art research in the context of real-life challenges.

The proposed IDAC enhances noise robustness in simulated and real-world noisy medical imaging experiments compared to conventional noise-robust DL methods. Compared to CE training, it boosts the performance by 5.1% for automatically labeled clinical in-house data. Simulated noise scenarios show a performance increase of up to 13.5%. The enhanced performance is evident across different use cases and various noise levels.

Beyond its enhanced performance, the IDAC is straightforward to implement, does not increase model complexity and shows a lower tendency to overfit, making it more user-friendly. Figure 8.1 represents the validation performance of a 30% noise simulation for the detection of pleural effusion and different abstaining weights α. It highlights the superiority of the IDAC loss compared to DAC and CE training based on performance and overfitting tendency. These results indicate the advantages of incorporating contextual information about the expected label noise. In addition, we evaluate the context-aware loss function and baseline methods on a real-life noisy clinical in-house data set, underlining the potential clinical impact of the proposed IDAC loss function. Consequently, the proposed context-aware IDAC can be a valuable tool for researchers, companies, and clinics aiming to develop accurate and reliable

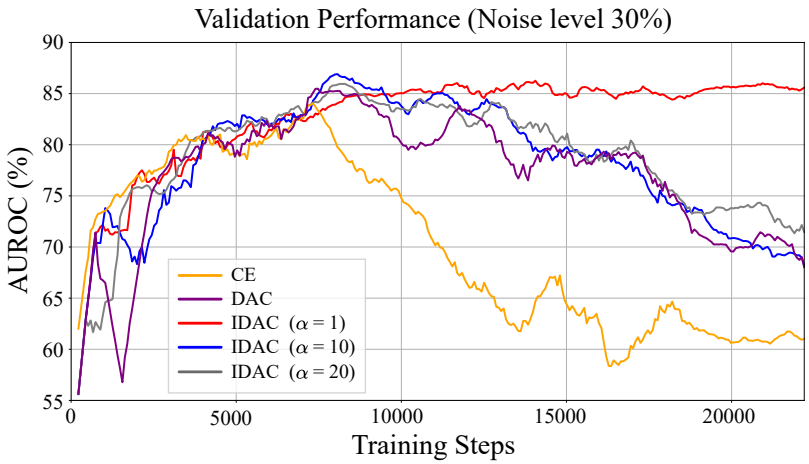


Figure 8.1: Smoothed validation performance of proposed IDAC training for classifying pleural effusion considering a simulated noise scenario. Different abstention weights α for a estimated noise levels of 30% are investigated and compared with CE and DAC training. An enhanced performance and mitigated overfitting tendencies are observed for the IDAC training. Specifically, the optimal performance is attained with $\alpha = 10$, whereas the least tendency for overfitting is observed at $\alpha = 1$.

DDSS from routine clinical data (with automatically generated labels).

In addition to the Chapters 3 and 5 this work emphasizes that the loss function is a suitable approach to incorporate contextual information into the DL training pipeline. While in Chapter 5 missing annotations are considered as prior label quality information, we prioritized label noise as contextual insight in this chapter. The given work therefore underlines that different types of contextual information about the label quality can be taken into account by context-aware loss functions.

8.2 Author's Contributions

My contributions to this paper as first author (shared first authorship) are as follows:
I was responsible for the research, planning and development of the context-aware loss function. I implemented the required DL pipelines and conducted the tuning of the models. Equally, together with my co-author S. Nowak, I generated the final tables and plots of the paper. We were equally responsible for the writing of the text, where my contribution relays most strongly on the related work, the introduction of the algorithm contributions and the result analysis. The paper was written with helpful support through discussions and proofreading from all co-authors.

CHAPTER 9

Conclusion

The concept of context-aware DL is based on the notion of developing data-driven solutions through the provision of contextual information. The area of medical image analysis lends itself to the application of such techniques due to the abundance of contextual information in place. In this thesis, novel context-aware DL methods were proposed to analysis medical image data leveraging the two contextual information types, expert knowledge and prior insights on the label quality of the utilized data set.

Specifically, we demonstrated that expert knowledge enhances the analysis of CXR images through context-aware DL methods, incorporating insights about the elements of bilateral symmetry of the lung fields. The given research presented, to the best of our knowledge, the initial investigation of the elements of bilateral symmetry of the lung fields in CXR data as potential contextual information for DL methods to detect lung diseases. We efficiently utilized symmetry-aware Siamese architectures and a symmetry-aware contrastive loss function to integrate the contextual information. Furthermore, we expanded context-aware DL to the analysis of lumbar MRI scans. We proposed a rule-based expert system calculating pertinent measurements that utilizes data-driven segmentation to facilitate interpretable DDSS. Our findings emphasized that expert knowledge can be effectively integrated through a two-step process. Overall we highlighted the potential of expert knowledge to improve performance, data-efficiency and interpretability of DL methods, critical aspects for the development of efficient DL-based DDSS.
Moreover, we investigated the potential of prior insights regarding the label quality for DL methods for various medical image analysis challenges. We demonstrated that context-aware loss functions are capable of efficiently managing missing labels. We considered which labels are unobserved as contextual information. Our proposed novel context-aware SPML loss functions outperformed state-of-the-art SPML methods for medical image analysis. In addition, we highlighted that employing context-aware DL methods can enhance performance when dealing with manually and automatically generated labels (i.e. gold and silver annotations). We proposed a context-aware pre-training strategy. Inferior annotations were solely used for pre-training, while we fine-tuned the model on gold-annotated samples. This context-aware strategy outperformed purely data-driven training, underlining the potential of the context-aware DL training process. Finally, we presented a novel context-aware loss function to process data sets with labels noise. The proposed method abstains potentially noisy samples during training by incorporating prior information about the expected label noise ratio of the given data set. We achieved enhanced noise-robustness and mitigated overfitting tendency, while

proposing an easy-to-implement solution without increased model complexity. Overall, the thesis highlights the significant benefits of context-aware DL by mitigating current challenges in the medical image analysis field. We therefore contributed to the development of context-aware DL-based DDSS that leverage clinical routine data, ultimately aiming to alleviate the workload of medical professionals and improving patient outcomes.

9.1 Discussion

This thesis addresses the following central research questions: 1) Can expert knowledge be leveraged to mitigate current challenges in medical image analysis, and if so, how? 2) Is it possible to utilize contextual information to attain good performance in the multi-label classification of medical image data despite a substantial proportion of missing labels, and if so, how? 3) Is it feasible to integrate contextual information about the label quality to enhance performance when handling data sets with label noise, and if so, how? In this section the achieved results are discussed in relation to each research question:

Can expert knowledge be leveraged to mitigate current challenges in medical image analysis, and if so, how?

Yes, it is possible to leverage expert knowledge to mitigate current challenges, such as interpretability or data-efficient training. This question was thoroughly addressed in Chapters 2 and 3, demonstrating expert knowledge as a valuable insight for the analysis of CXR images. In this work, context-aware DL methods were introduced, leveraging the elements of bilateral symmetry of lung fields in CXR scans. Specifically, symmetry-aware Siamese network architectures were proposed to incorporate this contextual information into the modeling process. Additionally, a symmetry-aware extension of a contrastive loss was introduced in Chapter 3 to enhance the integration of expert knowledge into the training pipeline. Experimental results suggested that these context-aware adaptations not only improved performance compared to purely data-driven baselines but also exhibited increased interpretability and data efficiency. To the best of our knowledge, this research represents *the first investigation of symmetry-aware DL based on the elements of bilateral symmetry of the lung fields in CXR scans* in order to detect lung diseases.

Furthermore, Chapter 4 explored this question, focusing on the automatic analysis of lumbar spine MRI scans. The objective was to calculate relevant distances for the diagnostic process automatically to enable anomaly detection. A rule-based expert knowledge system using DL-based segmentations of relevant vertebral bodies was presented, underlining that contextual information can be integrated in a two-stage process. Incorrect calculations were traceable. Chapter 4 further emphasized the potential of expert knowledge for the development of interpretable methods. Moreover, we demonstrated that context-aware DL is not only beneficial for the analysis of CXR data but is also highly relevant for other use cases, regardless of the image modality employed. In contrast to current lumbar MRI analysis research, we conducted our experiments based on a German patient cohort. The given research thus offered a highly pertinent evaluation of the proposed context-aware method using German medical image data.

Is it possible to utilize contextual information to attain good performance in the multi-label classification of medical image data despite a substantial proportion of missing labels, and if so, how?

Yes, we can leverage contextual information about the label quality, specifically missing labels, to achieve good performance despite only partially observed labels. This question was comprehensively investigated in Chapter 5, focusing on SPML training for medical image analysis. We proposed two novel context-aware SPML loss functions, incorporating the information regarding which labels are unobserved. For both functions, unobserved labels are mapped to negative labels during training. The aim of the proposed context-aware methods is to mitigate the influence of the wrongly mapped (noisy) labels. The introduced *novel, context-aware SPML loss functions* result in efficient implicit weighting factors during gradient calculations for unobserved labels, enhancing the noise-robustness of the DL method. Various experimental results underlined that these methods outperform state-of-the-art SPML strategies for the complex use case of medical image analysis. They achieved good performance with a significant portion of missing labels.
Additionally, the given research represents the first investigation of SPML training within the medical image analysis domain. Thus, connecting cutting-edge DL research with critical real-world challenges while offering a thorough evaluation of the application potential of SPML training.
We emphasized that contextual information regarding label quality is invaluable for the analysis of medical image data sets. In addition to Chapter 3, we highlighted that loss functions are particularly effective for incorporating contextual information during the training process. This holds true not only for expert knowledge as a type of contextual information but also for prior insights related to label quality.

Is it feasible to integrate contextual information about the label quality to enhance performance when handling data sets with label noise, and if so, how?

Yes, it is possible to achieve strong performance when dealing with data sets with label noise by employing context-aware methods. This research question is thoroughly discussed in the Chapters 6 and 7. The conducted experiments are based on clinical routine ICU CXR data provided by the University Hospital Bonn, leveraging manually (”gold”) and automatically (”silver”) generated labels. The annotations are based on the corresponding report of an image, using NLP methods for the automatic label generation. Automatic labeling enables a cost-effective and time-efficient generation of extensive annotated data sets, but label noise is introduced. In Chapter 6, we utilized a simpler rule-based labeler, while we employed a transformer-based NLP approach in Chapter 7, leading to higher label quality for the silver annotations. We proposed a context-aware pre-training strategy, first pre-training the network on silver-labeled data, and then fine-tuning it on gold samples. This method integrates the contextual information regarding the different label qualities of the gold and silver annotated sets during training. We demonstrated that *the context-aware inclusion of an extensive silver-labeled data set, can improve the performance of the model*, for both rule- and transformer based annotations.
In Chapter 7, we showed that this context-aware training can outperform conventional data-driven training (e.g. not differentiate between silver and gold annotations). Additionally, we demonstrated that context-aware pre-training can surpass transfer learning on public medical image data sets (for a limited volume of gold-annotated data) in Chapter 6. This highlights the potential of context-aware DL using prior insights regarding label quality to efficiently utilize silver-annotated data. However,

the given work outlines the need for noise-robust DL methods, specifically for medical image analysis. Please note that the given research presents the *initial processing of an ICU CXR in-house clinical routine data set provided by the University Hospital Bonn*. Consequently, it enhances the evaluation of context-aware DL methods for complex real-life clinical routine data.

Building on the insights from Chapter 7, we further explored the given research question in Chapter 8. We assumed that an estimate of the expected noise ratio is available. Since an assessment of the performance of automatic/human annotators on an independent test set with high label quality is crucial for managing the annotation process, this information is often provided. This chapter proposed a *novel, context-aware loss function* to leverage medical image data sets with noisy labels for the development of efficient DDSS.

The introduced loss function enables the abstaining of potential mislabeled samples during training. This abstention process is guided by an estimated measure of expected label noise ratio, which serves as crucial contextual information. By incorporating this noise ratio estimation, the model can more effectively distinguish between reliable and unreliable training examples, ultimately improving its generalization capabilities and performance on clean data.
Enhanced noise robustness is demonstrated not only through noise simulation experiments but also with clinical in-house data of ICU CXR scans utilizing automatically generated labels provided by the University Hospital Bonn. These results highlighted the potential clinical impact of the proposed context-aware loss function, enabling efficient noise-robust training for routine clinical data (with automatically generated labels). While improving noise robustness the introduced loss function additionally represents an easy-to-implement user-friendly DL method. Building on Chapter 3 and 5, this work underlined the potential of incorporating contextual information into the data-driven loss functions.

Further Contributions

This thesis not only addressed key research questions through novel algorithmic contributions but also tackled broader challenges in medical image analysis. Both proposed and state-of-the-art DL methods were evaluated with a focus on German patient cohorts. We utilized real-life noisy labels and automatically generated annotations to enhance the clinical relevance and impact of our investigations in crucial areas of medical image analysis.

Moreover, we introduced initial investigations of significant DL research areas, such as SPML training, within the context of medical imaging. Thus, we bridged the gap between current state-of-the-art DL research and the intricate field of medical image analysis. This integration not only advances the field but also lays the groundwork for more sophisticated and clinically impactful applications of these technologies in healthcare.

9.2 Outlook

While this thesis thoroughly addressed and answered the three primary research questions, new research findings lead to new research question. The following section briefly outlines potential future research directions that build upon our results, highlighting opportunities for further exploration and advancement in the field.

Enhancement of Context-aware DL Methods

This thesis presented powerful context-aware methods to analysis medical image data. Future work involves extending these methods to enhance the influence of contextual information on the training process. For instance, proposed loss functions designed to handle label noise resulting from incorrect or missing annotations could be extended by incorporating pseudo-label modules. The objective is to strengthen the generalization capability of these methods by providing potentially accurate annotations during the training process. Additionally, we consider a more comprehensive evaluation of the proposed methods as future work, including extended use cases and image modalities to further elucidate their potential clinical impact.

Integration of New Contextual Information

In this thesis, context-aware DL methods were proposed to leverage the two contextual information fields label quality and expert knowledge. However, several other forms of prior information are equally relevant in the analysis of medical images. For instance, imbalances in the training data set, particularly when using clinical routine data, frequently occur and can hinder model performance for underrepresented classes. The distribution of training data serves as valuable contextual information. Given our demonstration of the significant potential of context-aware losses the introduction of imbalance-aware loss functions emerges as a highly relevant area for future research.
Additionally, extensive contextual information based on the two discussed fields remains available for various use cases.

Inspired by Chapters 6 and 7, we may leverage patient demographic data to propose context-aware pre-training strategies. DICOM files, used for storing medical images, typically include patient metadata, and studies show that rough estimations of characteristics like age are feasible from CXR data [147]. This contextual information can therefore be exploited to develop cost-effective and time-efficient pre-training or multi-task learning approaches, potentially enhancing model performance.
Furthermore, we can adapt the concept of gold and silver generated labels to other use cases and image modalities. For instance, one can focus on the detection of pertinent entities on a pixel-based level for surgical video data. Silver annotations can be provided by interpolating in between annotated images. Context-aware loss functions or pre-training, discussed in this thesis, hold the potential to achieve remarkable performance by leveraging the generated extensive silver data set. The insights presented in this thesis regarding the integration of contextual information can inspire and support researchers in developing context-aware methods for new fields of contextual information.

Further Application

The proposed context-aware solution, while assessed within the medical image analysis domain, holds significant relevance for various applications. For instance, supervised classification systems utilized for land cover mapping require accurate reference databases, typically sourced from field measurements, thematic maps, or aerial photographs [148]. However, these data sets may contain label noise due to misregistration, update delays, or the inherent complexity of land cover [148]. Assessing the efficacy of our proposed context-aware DL solutions across these diverse application areas constitutes a crucial direction for future research. Additionally, manual annotation of multi-label

image data for complex use cases, such as road scene recognition for autonomous driving technologies [149], poses time and cost challenges. Investigating our proposed context-aware SPML loss functions for novel application fields presents a valuable future investigation. The insights shared in this thesis regarding the integration of contextual information can encourage and assist researchers in adapting the proposed context-aware methods to new application domains.

Context-aware DL and Foundation Models

The concept of foundation models revolves around the development of large-scale ML models trained on vast amounts of data, enabling them to generalize to tasks and data distributions beyond the employed training data [150]. While these models have achieved exceptional results in processing text data through Large Language Models such as ChatGPT-4 [151, 152], they are also being investigated for medical image analysis [38, 153, 154]. Despite their high performance in several use cases, these foundation models still face limitations such as data efficiency and imbalanced training data distributions [150]. Integrating contextual information into these models presents a potential solution, leading to more efficient models with better generalization. The exploration of how contextual information can be incorporated into foundation models represents highly relevant future work, aiming to develop DDSS with significant clinical impact based on foundation models.

Bibliography

[1] H. Chan et al. Deep learning in medical image analysis. In *Deep Learning in Medical Image Analysis: Challenges and Applications*, 2020.

[2] W. Zhu et al. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[3] X. Li et al. Multi-resolution convolutional networks for chest x-ray radiograph based lung nodule detection. In *Artificial Intelligence in Medicine*, 2020.

[4] X. Xie et al. A survey on incorporating domain knowledge into deep learning for medical image analysis. In *Medical Image Analysis*, 2021.

[5] S. Suganyadevi et al. A review on deep learning in medical image analysis. In *International Journal of Multimedia Information Retrieval*, 2022.

[6] J. Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

[7] X. Liu et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. In *The Lancet Digital Health*, 2019.

[8] A. Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. In *Nature*, 2017.

[9] Z. Salahuddin et al. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. In *Computers in Biology and Medicine*, 2022.

[10] H. Kim et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. In *The Lancet Digital Health*, 2020.

[11] P. Rajpurkar et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. In *arXiv preprint arXiv:1711.05225*, 2017.

[12] H. Schneider et al. Towards symmetry-aware pneumonia detection on chest x-rays. In *IEEE Symposium Series on Computational Intelligence*, 2022.

[13] H. Schneider et al. Symmetry-aware siamese network: Exploiting pathological asymmetry for chest x-ray analysis. In *International Conference on Artificial Neural Networks*, 2023.

[14] H. Schneider et al. Segmentation and analysis of lumbar spine mri scans for vertebral body measurements. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2023.

[15] H. Schneider et al. Is one label all you need? single positive multi-label training in medical image analysis. In *IEEE International Conference on Big Data*, 2023.

[16] H. Schneider et al. Improving intensive care chest x-ray classification by transfer learning and automatic label generation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2022.

[17] S. Nowak et al. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. In *European Radiology*, 2024.

[18] H. Schneider et al. Informed deep abstaining classifier: Investigating noise-robust training for diagnostic decision support system. In *International Conference on Neural Information Processing*, 2024.

[19] S. Hussain et al. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. In *BioMed Research International*, 2022.

[20] S. Siuly et al. Medical big data: neurological diseases diagnosis through medical data analysis. In *Data Science and Engineering*, 2016.

[21] M. Sermesant et al. Applications of artificial intelligence in cardiovascular imaging. In *Nature Reviews Cardiology*, 2021.

[22] M. Jafari et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. In *Computers in Biology and Medicine*, 2023.

[23] S. Bauer et al. A survey of mri-based medical image analysis for brain tumor studies. In *Physics in Medicine and Biology*, 2013.

[24] N. Sharma et al. Automated medical image segmentation techniques. In *Journal of Medical Physics*, 2010.

[25] M. Rana et al. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. In *Multimedia Tools and Applications*, 2023.

[26] S.K. Zhou et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. In *Proceedings of the IEEE*, 2021.

[27] P. Shah et al. Missed non–small cell lung cancer: radiographic findings of potentially resectable lesions evident only in retrospect. In *Radiology*, 2003.

[28] A. Del Ciello et al. Missed lung cancer: when, where, and why? In *Diagnostic and Interventional Radiology*, 2017.

[29] J. Iglehart. The new era of medical imaging—progress and pitfalls. In *New England Journal of Medicine*, 2006.

[30] J. Iglehart. Health insurers and medical-imaging policy: a work in progress. In *New England Journal of Medicine*, 2009.

[31] F. Ritter et al. Medical image analysis. In *IEEE Pulse*, 2011.

[32] D. Shen et al. Deep learning in medical image analysis. In *Annual Review of Biomedical Engineering*, 2017.

[33] Q.D. Buchlak et al. Effects of a comprehensive brain computed tomography deep learning model on radiologist detection accuracy. In *European Radiology*, 2024.

[34] S. Budd et al. A survey on active learning and human-in-the-loop deep learning for medical image analysis. In *Medical Image Analysis*, 2021.

[35] S. Candemir et al. A review on lung boundary detection in chest x-rays. In *International Journal of Computer Assisted Radiology and Surgery*, 2019.

[36] M. Li et al. Medical image analysis using deep learning algorithms. In *Frontiers in Public Health*, 2023.

[37] J. Lorkowski et al. Shortage of physicians: a critical review. In *Medical Research and Innovation*, 2021.

[38] J. Ma et al. Segment anything in medical images. In *Nature Communications*, 2024.

[39] Z. Hu et al. Deep learning for image-based cancer detection and diagnosis- a survey. In *Pattern Recognition*, 2018.

[40] P.K. Mall et al. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. In *Healthcare Analytics*, 2023.

[41] A. Mikolajczyk et al. Data augmentation for improving deep learning in image classification problem. In *International Interdisciplinary PhD Workshop (IIPhDW)*, 2018.

[42] G. Litjens et al. A survey on deep learning in medical image analysis. In *Medical Image Analysis*, 2017.

[43] D. Karimi et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. In *Medical Image Analysis*, 2020.

[44] Z. Zhang et al. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

[45] X. Zhou et al. Asymmetric loss functions for noise-tolerant learning: Theory and applications. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[46] X. Ye et al. Active negative loss functions for learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2023.

[47] B. Wu et al. Multi-label learning with missing labels. In *International Conference on Pattern Recognition*, 2014.

[48] Q. Teng et al. A survey on the interpretability of deep learning in medical diagnosis. In *Multimedia Systems*, 2022.

[49] Y. Hu et al. Adversarial deformation regularization for training image registration neural networks. In *Medical Image Computing and Computer Assisted Intervention*, 2018.

[50] Q. Guan et al. Thorax disease classification with attention guided convolutional neural network. In *Pattern Recognition Letters*, 2020.

[51] P. Chlap et al. A review of medical image data augmentation techniques for deep learning applications. In *Journal of Medical Imaging and Radiation Oncology*, 2021.

[52] M. Xie et al. Label-aware global consistency for multi-label learning with single positive labels. In *Advances in Neural Information Processing Systems*, 2022.

[53] X. Xie et al. Canet: Context aware network with dual-stream pyramid for medical image segmentation. In *Biomedical Signal Processing and Control*, 2023.

[54] Z. Huang et al. Context-aware legal citation recommendation using deep learning. In *International Conference on Artificial Intelligence and Law*, 2021.

[55] L. Miranda et al. A survey on the use of machine learning methods in context-aware middlewares for human activity recognition. In *Artificial Intelligence Review*, 2022.

[56] L. Pei et al. Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images. In *Scientific Reports*, 2020.

[57] Q. Zhu et al. A global context-aware and batch-independent network for road extraction from vhr satellite imagery. In *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021.

[58] Y. Ding et al. Multiscale graph sample and aggregate network with context-aware learning for hyperspectral image classification. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[59] D. Zhang et al. Self-supervised image denoising for real-world images with context-aware transformer. In *IEEE Access*, 2023.

[60] U. Fang et al. Robust image clustering via context-aware contrastive graph learning. In *Pattern Recognition*, 2023.

[61] B. Tu et al. A new context-aware framework for defending against adversarial attacks in hyperspectral image classification. In *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[62] A.R.H. Ali et al. Automating the abcd rule for melanoma detection: A survey. In *IEEE Access*, 2020.

[63] E. Cole et al. Multi-label learning from single positive labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[64] Z. Huang et al. Twin contrastive learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[65] Y. Li et al. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[66] X. Xia et al. Combating noisy labels with sample selection by mining high-discrepancy examples. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[67] X. Xing et al. Vision-language pseudo-labels for single-positive multi-label learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[68] H. Wei et al. Mitigating memorization of noisy labels by clipping the model prediction. In *International Conference on Machine Learning*, 2023.

[69] D. Zhou et al. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, 2022.

[70] G. Varoquaux et al. Machine learning for medical imaging: methodological failures and recommendations for the future. In *NPJ Digital Medicine*, 2022.

[71] S. Jaeger et al. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. In *Quantitative Imaging in Medicine and Surgery*, 2014.

[72] G.M.M. Alshmrani et al. A deep learning architecture for multi-class lung diseases classification using chest x-ray (cxr) images. In *Alexandria Engineering Journal*, 2023.

[73] W. Al-Dhabyani et al. Dataset of breast ultrasound images. In *Data in Brief*, 2020.

[74] N. Bien et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. In *PLoS Medicine*, 2018.

[75] S.M. Niehues et al. Deep-learning-based diagnosis of bedside chest x-ray in intensive care and emergency medicine. In *Investigative Radiology*, 2021.

[76] V.K. Raghu et al. Deep learning to estimate biological age from chest radiographs. In *Cardiovascular Imaging*, 2021.

[77] M. Gross et al. Automated mri liver segmentation for anatomical segmentation, liver volumetry, and the extraction of radiomics. In *European Radiology*, 2024.

[78] M. Mohri et al. *Foundations of machine learning*. MIT press, 2018.

[79] C.M. Bishop et al. *Pattern recognition and machine learning*. Springer, 2006.

[80] M.M. Elsayed et al. Dialysis recovery time: associated factors and its association with quality of life of hemodialysis patients. In *BMC Nephrology*, 2022.

[81] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, 1991.

[82] S.J. Russell et al. *Artificial intelligence: a modern approach*. Pearson, 2016.

[83] U. Von Luxburg et al. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, 2011.

[84] J.R. Quinlan. Induction of decision trees. In *Machine Learning*, 1986.

[85] P.B. Schiilkop et al. Extracting support data for a given task. In *First International Conference on Knowledge Discovery and Data Mining*, 1995.

[86] Y. Lecun et al. Pattern recognition and neural networks. In *The Handbook of Brain Theory and Neural Networks*, 1995.

[87] I. Goodfellow et al. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[88] M. Belkin et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. In *Proceedings of the National Academy of Sciences*, 2019.

[89] Y. LeCun et al. Deep learning. In *Nature*, 2015.

[90] Y. LeCun et al. A theoretical framework for back-propagation. In *Connectionist Models Summer School*, 1988.

[91] D.P. Kingma. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014.

[92] G.E. Hinton et al. A fast learning algorithm for deep belief nets. In *Neural Computation*, 2006.

[93] M.A. Morid et al. A scoping review of transfer learning research on medical image analysis using imagenet. In *Computers in Biology and Medicine*, 2021.

[94] S. Niu et al. A decade survey of transfer learning (2010–2020). In *IEEE Transactions on Artificial Intelligence*, 2020.

[95] K. He et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*, 2015.

[96] K. Weiss et al. A survey of transfer learning. In *Journal of Big Data*, 2016.

[97] J. Nam et al. Large-scale multi-label text classification—revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, 2014.

[98] Q. Wang et al. A comprehensive survey of loss functions in machine learning. In *Annals of Data Science*, 2020.

[99] D. Rumelhart et al. Learning representations by back-propagating errors. In *Nature*, 1986.

[100] Y. LeCun et al. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, 1995.

[101] Y. Zhou et al. Computation of optical flow using a neural network. In *Neural Networks*, 1988.

[102] G. Huang et al. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[103] K. He et al. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[104] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted Intervention*, 2015.

[105] R.R. Selvaraju et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In *International Journal of Computer Vision*, 2017.

[106] A. Chattopadhay et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[107] I. Melekhov et al. Siamese network features for image matching. In *International Conference on Pattern Recognition*, 2016.

[108] J. Bromley et al. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, 1993.

[109] R. Hadsell et al. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[110] Z. Alaverdyan et al. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. In *Medical Image Analysis*, 2020.

[111] A. Hekler et al. Effects of label noise on deep learning-based skin cancer classification. In *Frontiers in Medicine*, 2020.

[112] A. Basu et al. Covid-19 detection from ct scans using a two-stage framework. In *Expert Systems with Applications*, 2022.

[113] S. Wang et al. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In *IEEE International Symposium on Biomedical Imaging*, 2017.

[114] I. Carbone et al. *Thoracic Radiology*. Springer, 2020.

[115] P. Mildenberger et al. Introduction to the dicom standard. In *European Radiology*, 2002.

[116] M. Reiser et al. *Duale Reihe*. Thieme, 2011.

[117] A. Ganapathy et al. Routine chest x-rays in intensive care units: a systematic review and meta-analysis. In *Critical Care*, 2012.

[118] S. Raoof et al. Interpretation of plain chest roentgenogram. In *Chest*, 2012.

[119] B.A. Winegar et al. Magnetic resonance imaging of the spine. In *Polish Journal of Radiology*, 2020.

[120] J.L. Chazen et al. Rapid lumbar mri protocol using 3d imaging and deep learning reconstruction. In *Skeletal Radiology*, 2023.

[121] J. Cheung et al. Verification of measurements of lumbar spinal dimensions in t1-and t2-weighted magnetic resonance imaging sequences. In *The Spine Journal*, 2014.

[122] J Van Goethem, P Parizel, and J Jinkins. Mri of the postoperative lumbar spine. In *Neuroradiology*, 2002.

[123] M.C. Fu et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. In *The Spine Journal*, 2014.

[124] A. Suri et al. Vertebral deformity measurements at mri, ct, and radiography using deep learning. In *Radiology: Artificial Intelligence*, 2021.

[125] E. Çallı et al. Deep learning for chest x-ray analysis: A survey. In *Medical Image Analysis*, 2021.

[126] E. Kotei et al. A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest x-ray images. In *Archives of Computational Methods in Engineering*, 2024.

[127] L. Cui et al. Deep symmetric three-dimensional convolutional neural networks for identifying acute ischemic stroke via diffusion-weighted images. In *Journal of X-ray Science and Technology*, 2021.

[128] A. Barman et al. Determining ischemic stroke from ct-angiography imaging using symmetry-sensitive convolutional networks. In *IEEE International Symposium on Biomedical Imaging*, 2019.

[129] Y. Liu et al. From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network. In *Medical Image Computing and Computer Assisted Intervention*, 2019.

[130] C. Germann et al. Performance of a deep convolutional neural network for mri-based vertebral body measurements and insufficiency fracture detection. In *European Radiology*, 2023.

[131] R.F. Masood et al. Deep learning based vertebral body segmentation with extraction of spinal measurements and disorder disease classification. In *Biomedical Signal Processing and Control*, 2022.

[132] S. Pang et al. Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. In *IEEE Transactions on Medical Imaging*, 2020.

[133] D. Zhou et al. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, 2022.

[134] S. Nowak et al. Transformer-based structuring of free-text radiology report databases. In *European Radiology*, 2023.

[135] H.H. Pham et al. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. In *Neurocomputing*, 2021.

[136] S.M. Niehues et al. Deep-learning-based diagnosis of bedside chest x-ray in intensive care and emergency medicine. In *Investigative Radiology*, 2021.

[137] K. Ding et al. Improve noise tolerance of robust loss via noise-awareness. In *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[138] S. Thulasidasan et al. Combating label noise in deep learning using abstention. In *International Conference on Machine Learning*, 2019.

[139] X. Ma. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, 2020.

[140] W. Kusakunniran et al. Covid-19 detection and heatmap generation in chest x-ray images. In *Journal of Medical Imaging*, 2021.

[141] K.C. Santosh et al. Automated chest x-ray screening: Can lung region symmetry help detect pulmonary abnormalities? In *IEEE Transactions on Medical Imaging*, 2017.

[142] Ž Vujović et al. Classification model evaluation metrics. In *International Journal of Advanced Computer Science and Applications*, 2021.

[143] H. Chen et al. Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images. In *European Conference Computer Vision European Conference*, 2020.

[144] K. He et al. Mask r-cnn. In *IEEE International Conference on Computer Vision*, 2017.

[145] J. Bertels et al. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention*, 2019.

[146] M. Zhang et al. A review on multi-label learning algorithms. In *IEEE Transactions on Knowledge and Data Engineering*, 2013.

[147] P.H. Yi et al. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. In *Emergency Radiology*, 2021.

[148] C. Pelletier et al. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. In *Remote Sensing*, 2017.

[149] L. Chen et al. Deep integration: A multi-label architecture for road scene recognition. In *IEEE Transactions on Image Processing*, 2019.

[150] B. Azad et al. Foundational models in medical imaging: A comprehensive survey and future vision. In *arXiv preprint arXiv:2310.18689*, 2023.

[151] Y. Chang et al. A survey on evaluation of large language models. In *ACM Transactions on Intelligent Systems and Technology*, 2024.

[152] J. Achiam et al. Gpt-4 technical report. In *arXiv preprint arXiv:2303.08774*, 2023.

[153] J. Cheng et al. Sam-med2d. In *arXiv preprint arXiv:2308.16184*, 2023.

[154] S. Pai et al. Foundation model for cancer imaging biomarkers. In *Nature Machine Intelligence*, 2024.

List of Figures