Computer Methods and Programs in Biomedicine 221 (2022) 106874

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Medical deep learning—A systematic meta-review

Jan Egger[a,b,c,d,e,*], Christina Gsaxner[a,b,c], Antonio Pepe[a,c], Kelsey L. Pomykala[d], Frederic Jonske[c,d], Manuel Kurz[a,c], Jianning Li[a,c,d], Jens Kleesiek[d,e,f]

[a] Institute of Computer Graphics and Vision, Faculty of Computer Science and Biomedical Engineering, Graz University of Technology, Inffeldgasse 16, 8010 Graz, Styria, Austria
[b] Department of Oral &Maxillofacial Surgery, Medical University of Graz, Auenbruggerplatz 5/1, 8036 Graz, Styria, Austria
[c] Computer Algorithms for Medicine Laboratory, Graz, Styria, Austria
[d] Institute for AI in Medicine (IKIM), University Medicine Essen, Girardetstraße 2, 45131 Essen, Germany
[e] Cancer Research Center Cologne Essen (CCCE), University Medicine Essen, Hufelandstraße 55, 45147 Essen, Germany
[f] German Cancer Consortium (DKTK), Partner Site Essen, Hufelandstraße 55, 45147 Essen, Germany

ARTICLE INFO

Article history:
Received 5 January 2021
Revised 22 April 2022
Accepted 10 May 2022

Keywords:
Deep learning
Artificial neural networks
Machine learning
Data analysis
Image analysis
Medical image analysis
Medical image processing
Medical imaging
Patient data
Pathology
Detection
Segmentation
Registration
Generative adversarial networks
PubMed
Systematic
Review
Survey
Meta-review
Meta-survey

ABSTRACT

Deep learning has remarkably impacted several different scientific disciplines over the last few years. For example, in image processing and analysis, deep learning algorithms were able to outperform other cutting-edge methods. Additionally, deep learning has delivered state-of-the-art results in tasks like autonomous driving, outclassing previous attempts. There are even instances where deep learning outperformed humans, for example with object recognition and gaming. Deep learning is also showing vast potential in the medical domain. With the collection of large quantities of patient records and data, and a trend towards personalized treatments, there is a great need for automated and reliable processing and analysis of health information. Patient data is not only collected in clinical centers, like hospitals and private practices, but also by mobile healthcare apps or online websites. The abundance of collected patient data and the recent growth in the deep learning field has resulted in a large increase in research efforts. In Q2/2020, the search engine PubMed returned already over 11,000 results for the search term 'deep learning', and around 90% of these publications are from the last three years. However, even though PubMed represents the largest search engine in the medical field, it does not cover all medical-related publications. Hence, a complete overview of the field of 'medical deep learning' is almost impossible to obtain and acquiring a full overview of medical sub-fields is becoming increasingly more difficult. Nevertheless, several review and survey articles about medical deep learning have been published within the last few years. They focus, in general, on specific medical scenarios, like the analysis of medical images containing specific pathologies. With these surveys as a foundation, the aim of this article is to provide the first high-level, systematic meta-review of medical deep learning surveys.

1. Introduction

Deep learning [1] had a remarkable impact on different scientific fields during the last years. This was demonstrated in numerous tasks, where deep learning approaches were able to outperform the standard methods, including image processing and analysis [2,3]. Moreover, deep learning delivers reasonable results in tasks that could not have been performed automatically before, like autonomous driving [4,5]. There are even applications where deep learning outperformed humans, like in object recognition [6] or games [7,8].

A field in which this development has begun to show huge potential is the medical domain. With the collection of large quantities of patient records and data, and a trend towards personalized treatments, there is a great need for automatic and reliable processing and analysis of this information [9]. Patient data is not only collected in clinical centers, like hospitals and private practices, but also by mobile healthcare apps or online websites. Together this resulted in new, massive research efforts during the last years. In Q2 of 2020, the search engine PubMed returns already over 11.000 results for the search term "*deep learning*", and around

* Corresponding author.
E-mail address: egger@tugraz.at (J. Egger).

https://doi.org/10.1016/j.cmpb.2022.106874

90% of these publications are from the last three years. However, even though PubMed represents the largest search engine in the medical domain, it does not cover all medical-related publications. For example, medical topics are also covered in primary venues for computer science research, most often conferences [10,11]. Despite their high impact and consideration within the community, conference proceedings are usually not listed under PubMed, with only a few exceptions, like the prestigious annual conference '*Medical Image Computing and Computer Assisted Intervention*' (MICCAI). In addition, there are rather technical, non-interdisciplinary conferences, for example in computer vision, through which very influential research on medical applications is published [12]. These contributions are often overlooked by medical search engines. This does not relate to survey and review articles, which, due to their length, are generally published in peer-reviewed, PubMed-indexed journals. However, for this reason, it is possible for a review article to miss some relevant contributions.

Taking all these considerations into account, a complete overview of the field of medical deep learning is almost impossible to obtain and acquiring a full overview of medical sub-fields becomes increasingly more difficult. Nevertheless, several review and survey articles about medical deep learning have been published within the last years. They focus, in general, on specific medical scenarios, such as the analysis of certain medical images containing specific pathologies, like the automatic detection of a cardiovascular disorder in computed tomography angiography acquisitions [13]. In this context, the aim of this contribution is to provide an introductory, high-level and systematic meta-review of medical deep learning surveys. Modeled after existing meta-reviews in the medical domain, such as the systematic review of systematic reviews of homeopathy [14], or the survey of surveys on the use of visualization for interpreting machine learning models [15] in a technical domain. The authors are not aware of any meta-review in medical deep learning or general deep learning so far. Compared to medicine, which has a millennia-old tradition, computer science is a very young discipline. Nonetheless, if this discipline continues growing at the current pace, meta-reviews like this will become more and more common.

In this publication, we present all review and survey articles published from 2017 to 2019 found by a systematic PubMed search (see Search Strategy). We did not include articles published after 2019, since we also present the citations of the reviewed publications. Thus, the relatively new reviews from 2020 are still 'undercited' in comparison to the reviews from the previous years (as can also be seen in the decreasing number of citations for 2017: 6089 citations, 2018: 947 citations and 2019: 408 citations), and one aim of this contribution is to give an overall impression of the impact these works have already had on their respective scientific fields. Table 1 gives an overview of the number of reviews published each year, from 2017 to 2019. Furthermore, the table shows the sum of the overall references and citations for each year according to Google Scholar (status as of August 2020). Tables 2–4 describe the publications of each year in more detail.

Systematic literature review phase. For our systematic review, we started with planning the overall structure and main headings of this manuscript, orienting on existing surveys and meta-surveys in the literature. Next, we decided on the databases and years of publication that we wanted to include in our meta-survey. While keeping in mind the overall number of publications we want to cover within our manuscript. Subsequently, we performed the final literature search (see next paragraph *Search Strategy*), summarized every survey and extracted the citations, main architectures, evaluations, pros/cons, challenges and future directions. Finally, we analyzed the commonalities and drew a conclusion resulting in a discussion and future outlook.

Search Strategy. For this systematic meta-review, a search in PubMed for the keyword 'Deep Learning' together with any keyword including {'Review', 'Survey'} was performed. Based on the titles and abstracts, all records which were not actually review or survey contributions in the medical field, like [16,17] and [18], or were not written in English, like [19,20], or are veterinarian reviews [21], or are about a human learning strategy called *Deep Learning* [22], were excluded (while the term "*deep learning*" was coined by Geoffrey Hinton in terms of learning deep neural networks in 2006 [23,24], the term seemed to have existed much earlier in educational psychology [25]. Note further, that non-shallow neural networks had already become an explicit research subject by the early 1990s, when they also became practically feasible to some extent through the help of unsupervised learning [26]). This ultimately resulted in a total number of 43 review or survey publications about deep learning in the medical field, which are covered within this systematic meta-review. Summarized, this high-level systematic meta-review provides an overview of the published medical deep learning reviews and surveys in PubMed, as well as their references and citations (status as of August 2020). Note that our systematic search strategy does not cover all topics in medical deep learning, like a survey about uncertainty quantification in deep learning applications in medical data analysis [27]. However, we did not want to "break" our systematic meta-review search by adding what is arguably arbitrary additional literature.

Manuscript Outline. The main body of this contribution presents exclusively reviews and surveys on medical deep learning from a systematic PubMed search. To keep the manuscript concise for the reader, we provide only high-level summaries and excerpts, mainly form the review and survey abstracts (note that some of the presented reviews cover up to several hundred publications themselves). Thus, every review or survey publication will be summarized in around 100 to 200 words. However, by pointing to the associated publications via the keyword classifications and chronological arrangement of the presented medical deep learning reviews or surveys, the interested reader should be able to dive deeper into the specific categories and sub-categories. The rest of this manuscript is organized as follows: Section 2 presents the overview of the medical deep learning reviews and surveys divided into the years of publications from 2017 to 2019 in chronological order, beginning with the first published work in the respective year. The final Section 3 concludes this contribution with a discussion and outlines areas of future directions.

Research questions. The overall aim of this systematic meta-review is to analyze reviews and surveys published between 2017 and 2019 in medical deep learning. In doing so, we defined the following main research questions for our study:

Table 1
Overview of published reviews of deep learning in the medical field from beginning 2017 to end of 2019 according to PubMed and number of citations according to Google Scholar (status as of August 2020).

Year ▼	Number of publications	Number of references	Citations (until August 2020)
2017	7	1060	6089
2018	15	1684	947
2019	21	2279	408
Sum	**43**	**5023**	**7444**

Table 2
List of published reviews of deep learning in the medical field in 2017 according to PubMed and number of citations according to Google Scholar (status as of August 2020); ordered by epub (electronic publication) date.

Medical field/subject	Publications	Date (epub) ▼	Number of references	Citations (until August 2020)
Medical image analysis (I.)	Shen et al. [28]	March 09, 2017	117	1232
Healthcare	Miotto et al. [29]	May 06, 2017	119	624
Medical image analysis (II.)	Litjens et al. [30]	Jul. 26, 2017	439	3696
Stroke management	Feng et al. [31]	Sep. 27, 2017	55	40
Analysis of molecular images in cancer	Xue et al. [32]	Oct. 15, 2017	60	23
Health-record analysis	Shickel et al. [33]	Oct. 26, 2017	63	377
Microscopy image analysis	Xing et al. [34]	Nov. 22, 2017	207	97
Sum	**–**		**1060**	**6089**

Table 3
List of published reviews of deep learning in the medical field in 2018 according to PubMed and number of citations according to Google Scholar (status as of August 2020); ordered by epub (electronic publication) date.

Medical field/subject	Publications	Date (epub) ▼	Number of references	Citations (until August 2020)
Toxicity of chemicals	Tang et al. [35]	Mar. 01, 2018	103	13
Pulmonary nodule diagnosis	Yang et al. [36]	Apr. 2018	42	14
Physiological signals	Faust et al. [37]	Apr. 11, 2018	166	301
DNA sequencing	Celesti et al. [38]	Apr. 12, 2018	52	14
Radiotherapy	Meyer et al. [39]	May 17, 2018	234	86
Ophthalmology	Grewal et al. [40]	May 30, 2018	33	29
Electronic health records	Xiao et al. [41]	Jun. 08, 2018	123	146
Bioinformatics	Lan et al. [42]	Jun. 28, 2018	127	85
Personalized medicine	Zhang et al. [43]	Aug. 07, 2018	142	8
1-D biosignals	Ganapathy et al. [44]	Aug. 29, 2018	117	19
Omics	Zhang et al. [45]	Sep. 26, 2018	143	40
Sport-specific movement recognition	Cust et al. [46]	Oct. 11, 2018	98	35
Diabetic retinopathy	Nielsen et al. [47]	Nov. 03, 2018	42	12
Image cytometry	Gupta et al. [48]	Dec. 19, 2018	137	46
Radiology	Mazurowski et al. [49]	Dec. 21, 2018	125	99
Sum	**–**		**1684**	**947**

Table 4
List of published reviews of deep learning in the medical field in 2019 according to PubMed and number of citations according to Google Scholar (status as of August 2020); ordered by epub (electronic publication) date.

Medical field/subject	Publications	Date (epub) ▼	Number of references	Citations (until August 2020)
Medical imaging	Biswas et al. [50]	Jan. 01, 2019	94	28
Brain cancer classification	Tandel et al. [51]	Jan. 18, 2019	123	33
Electroencephalogram	Craik et al. [52]	Feb. 26, 2019	123	91
Pulmonary nodule detection	Pehrson et al. [53]	Mar. 07, 2019	48	21
Neuro-oncology	Shaver et al. [54]	Jun. 14, 2019	81	9
Diabetic retinopathy	Asiri et al. [55]	Aug. 07, 2019	138	21
Cardiac arrhythmia	Parvaneh et al. [56]	Aug. 08, 2019	20	4
Protein structure	Wardah et al. [57]	Aug. 12, 2019	72	7
Electroencephalography	Roy et al. [58]	Aug. 14, 2019	249	101
Neurology	Valliani et al. [59]	Aug. 21, 2019	83	8
Cancer diagnosis	Munir et al. [60]	Aug. 23, 2019	167	20
Ultrasound	Akkus et al. [61]	Sep. 03, 2019	78	7
Radiation oncology	Boldrini et al. [62]	Oct. 01, 2019	64	10
Drug–drug interaction	Zhang et al. [63]	Nov. 04, 2019	180	4
Urology	Suarez-Ibarrola et al. [64]	Nov. 05, 2019	56	10
Sleep apnea	Mostafa et al. [65]	Nov. 12, 2019	93	5
Ophthalmic diagnosis	Sengupta et al. [66]	Nov. 22, 2019	123	13
Alzheimer's disease	Ebrahimighahnavieh et al. [67]	Nov. 27, 2019	201	4
Pulmonary nodule detection	Li et al. [68]	Nov. 29, 2019	60	3
Liver masses	Azer [69]	Dec. 15, 2019	45	5
Pulmonary medical imaging	Ma et al. [70]	Dec. 16, 2019	181	4
Sum	**–**		**2279**	**408**

1) What are the different applications of deep learning in medicine?
2) What are the methods most frequently or successfully employed by deep learning in medicine?
3) What are the strengths and limitations of these methods, especially with respect to the field they are applied to?
4) What are the key research gaps that are being investigated or should be investigated according to researchers?

2. Medical deep learning: a compact overview of reviews and surveys

This section presents an overview of review and survey publications in medical deep learning. The publications are arranged in three sub-sections by their year of publication, from 2017 to 2019. Within the yearly sub-sections, the publications are arranged chronologically by their date of publication starting with the first

Fig. 1. Collage mapping all figures from the reviewed articles to a left hemisphere brain surface.

published work in the corresponding year. Typically, review or survey contributions order the reviewed publications in categories, like medical image classification, object detection, segmentation, registration, and other tasks. However, for this meta-review we decided explicitly for an order by publication date to show the historical sequence in which they occurred to the reader. Still, the tables provide also a quick overview of the different categories. Hence, at the beginning of every section (2017, 2018 and 2019), the areas of the reviews are summarized in a listing, which corresponds to the chronical order of the publications of this year in the following descriptions. Consequently, Tables 1–4 are also divided into the years 2017 to 2019 and chronically ordered. Note, that the reviewed surveys can focus on a specific subject, like the survey about diabetic retinopathy screening, or span over a general field, like the survey about healthcare. Moreover, the tables present the number of referenced works and the current citations for every year and publication according to Google Scholar, which reveals an overall number of 5023 referenced works in the proposed reviews, and an overall number of 7444 citations for the reviews themselves (status as of August 2020). Furthermore, Fig. 1 shows a collage, where we mapped all figures of the reviewed articles to the surface of the left hemisphere of the brain. Finally, and equivalent to [37], Fig. 2 shows a network visualization for the review articles supplied keywords from 2017 to 2019. More specifically, the figure shows the co-occurrence network and the topic clusters for the article keywords, and it reveals the two main clusters, namely “humans” and “deep learning”, and their connections. Further main clusters center around the keywords “machine learning”, “neural networks (computer)” and “algorithms”. Overall, the clusters and connections show how the medical domain has been affected by deep learning in these years, covering a broad range of topics and applications.

2.1. Medical deep learning reviews in 2017

With the described search strategy, seven medical deep learning surveys published in 2017 were discovered. Fig. 3 shows a network visualization of the review article keywords from 2017 revealing the keyword “humans” with its connections as the main cluster. Further main keyword clusters are “deep learning” and “neural networks (computer)”, which also reveal the main commonalities and trends for the surveys in 2017. More specific topics in the surveys of 2017 are “electronic health records” and “diagnostic imaging”. The reaming clusters are of a more general nature, like “machine learning”, “algorithms” and “image processing”. The presented reviews from 2017 cite 1060 contributions and have been cited 6089 times (status as of August 2020). They cover the following categories and are ordered by *epub* (electronic publication) date in 2017 (see Table 2):

- Medical image analysis (I.);
- Healthcare;
- Medical image analysis (II.);
- Stroke management;
- Analysis of molecular images in cancer;
- Health-record analysis;
- Microscopy image analysis.

Medical image analysis (I.) – The aim of medical image analysis is to automatically or semi-automatically extract information from patient data. For instance, this could be an automatic determination of the tumor volume from a patient’s magnetic resonance imaging (MRI) scan with the aim to choose the appropriate therapy strategy. Shen et al. [28] introduce in their publication the basics of deep learning-based approaches and survey their suc-

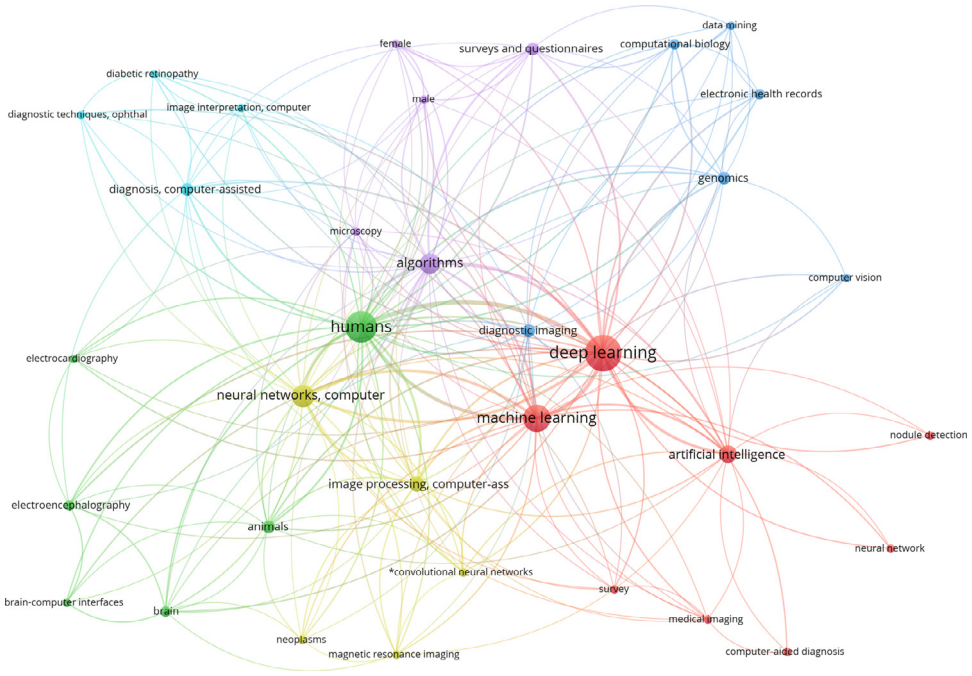

Fig. 2. Network visualization for the review articles supplied keywords from 2017 to 2019 performed with VOSviewer.

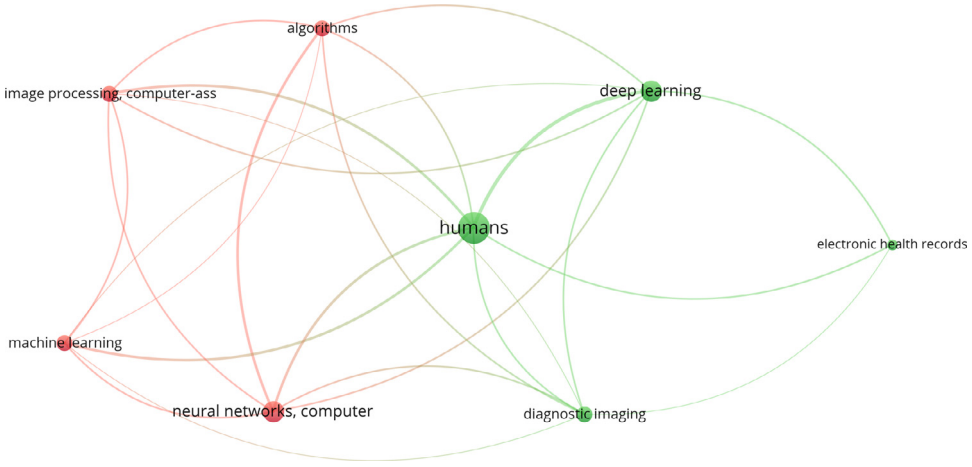

Fig. 3. Network visualization for the review articles supplied keywords in 2017 performed with VOSviewer.

cess in fields like image registration, tissue segmentation, anatomical/cell structures detection, computer-aided disease diagnosis, but also computer-aided disease prognosis. They conclude their work by pointing out remaining research challenges and give suggestions for future research directions that could advance medical image analysis.

Healthcare – The umbrella term healthcare envelopes the maintenance and advancement of people's health by diagnosis, prevention, treatment, but also recovery or even cure of illness, disease, injury, or any further physical or mental maladies. In that context, the survey article of Miotto et al. [29], reviews published research using deep learning-based approaches and technologies to improve the healthcare field. Centered on the analyzed publications, they conclude and propose that deep learning-based methods can be used to advance human health by exploring and exploiting big biomedical data. Furthermore, they depict limitations and the need for improved methods and applications and discuss future challenges in this area.

Medical image analysis (II.) – The publication of Litjens et al. [30] surveys the main deep learning-based concepts that are relevant for the area of medical image analysis. They summarize over 300 works within the area and analyze the usage of deep learning-based methods for object detection, image classification, segmentation, but also registration and further tasks. Moreover, they give compact, categorized outlines of studies in different areas of application, namely digital pathology, neurological, pulmonary, retinal, breast, as well as abdominal, cardiac, and musculoskeletal imaging. Finally, they give a summary of the current works at that time and discuss the remaining research questions and directions for upcoming research contributions.

Stroke management – Stroke can cause a long-term disability and a vast amount of research has been focused on using neuroimaging to explore regions of ischemia, which have not been affected by cellular death. In this context, Feng et al. [31] review clinical applications for deep learning-guided stroke management. They identify the following core topics for translating deep learning-based methods in the management of strokes, namely image segmentation, multimodal prognostication, but also radiomics (automated featurization).

Analysis of molecular images in cancer – Molecular imaging is of major interest for early cancer detection, because it opens the possibility to visualize biological changes on a molecular, but also on a

cellular level, which enables a quantitative analysis of them. Hence, Xue et al. [32] published a survey about deep learning-based applications for an automated analysis of molecular cancer image acquisitions. They survey the deep learning-based applications in the field of molecular imaging with regards to a segmentation of tumor lesions, classification of tumors, and a prediction of patient survival.

Health-record analysis – Health-record analysis explores the digital information stored in electronic health databases. The initial intention for storing information of patients are administrative tasks in healthcare, such as billing. However, subsequently health records also became interesting for numerous applications in clinical informatics for researchers. Hence, Shickel et al. [33] perform a review about deep learning-based research for clinical applications that depend on the analysis of health-record data. They explore numerous deep learning-based frameworks and techniques that have been used for various clinical tasks; for example, information extraction, representation learning, outcome prediction, phenotyping, and de-identification. The authors discovered several remaining research challenges, such as heterogeneity of data, the lack of available universal benchmark tests, and the interpretability of models. They finalize their analysis by recapitulating the recent works, as well as pointing out directions that could be upcoming research topics in deep learning-based processing of health-records.

Microscopy image analysis – Microscopy images are images acquired from a microscope that can be utilized for the characterization of various diseases, such as brain tumors, breast cancer or lung cancer. Xing et al. [34] explore the image analysis domain for medical microscopy by providing at first a dense overview of common deep neural networks. Then, they analyze and review state-of-the-art results of deep learning in the analysis of microscopy images, for example in the tasks of image segmentation, object detection and classification. The authors also describe several architectures in deep learning, namely convolutional and fully convolutional neural networks, but also deep belief and recurrent neural networks, and lastly, stacked autoencoders. Thereby, they investigate and depict the specific network structures for the different applications in the analysis of microscopy images. The authors end their review by outlining remaining research needs, and by highlighting possible research directions in the domain of deep learning-based processing of microscopy images.

2.1.1. Diving deeper: architectures, evaluations, pros, cons, challenges and future directions in 2017

Table 5 presents more details about the presented methods, pros, cons, evaluations, challenges and future directions for the reviews from the year 2017. All reported surveys share a number of important conclusions. They agree that deep learning is a promising approach for a wide variety of medical fields and tasks and predict that it will find increasing use in diagnosis, predictions, decision making and task automation. The deep learning-based methods explored by the respective surveys typically outperform previous state-of-the-art algorithms based on more naive approaches. In addition, the authors of the surveys all share the opinion that several challenges remain unsolved so far and will require additional exploration in the future. Among those are the inherently low explainability of deep learning approaches (often termed the "black box" problem) and lack of structured and expert-labeled or -annotated data, suggesting the creation of large-scale public datasets.

2.2. Medical deep learning reviews in 2018

With the proposed search strategy, 15 surveys were identified in medical deep learning from 2018. Fig. 4 shows a network visualization for the review articles supplied keywords in 2018 that reveals, equivalent to the surveys from 2017, the keywords "humans" and "deep learning", and its connections, as the main clusters. Further main keyword clusters center around the more general keywords "machine learning", "neural networks (computer)" and "algorithms". However, the smaller clusters around the keywords "electrocardiography", "computational biology", "surveys and questionnaires", "genomics" and "animals", show that the works in medical deep learning broadened in 2018 compared to 2017. The proposed reviews from 2018 themselves refer to 1684 contributions and have been cited 947 times (status as of August 2020). They cover the following categories, ordered by epub date in 2018 (Table 3):

- Toxicity of chemicals;
- Pulmonary nodule diagnosis;
- Physiological signals;
- DNA sequencing;
- Radiotherapy;
- Ophthalmology;
- Electronic health records;
- Bioinformatics;
- Personalized medicine;
- 1-D biosignals;
- Omics;
- Sport-specific movement recognition;
- Diabetic retinopathy;
- Image cytometry;
- Radiology.

Toxicity of chemicals – Toxicity testing and evaluation of chemicals is important for humans and animals, because they are exposed lifelong to natural and synthetic chemicals. Tang et al. [35] analyze in their work how deep learning-based tools can be a utilized for toxicity prediction, by building models for quantitative structure-activity relationships. They focus on large datasets, where classic data analysis techniques cannot deliver fast results. First, a technical overview about deep neural networks is provided by the authors. Then, recent works for the prediction of chemical toxicity models based on deep neural network approaches are explored. Finally, the important data sources for toxicity are outlined, remaining challenges are highlighted, and future directions for deep neural network-based approaches for the prediction of chemical toxicity are provided.

Pulmonary nodule diagnosis – A pulmonary nodule is a small, rounded opacity within the pulmonary interstitium. In their review, Yang et al. [36] present deep learning works that aid the decision-making in pulmonary nodule diagnosis. The deep learning-based methods they survey focus on computer-assisted feature extraction, false-positive reduction and nodule detection, but also on a benign-malignant classification in large volume scans of the chest.

Physiological signals – Physiological signals are signals from psycho-physiological measurements. In their survey, Faust et al. [37] review deep learning-based approaches utilized in healthcare applications that exploit physiological signals. Their bibliometric review revealed that the analyzed contributions focused mainly on Electromyograms (EMGs), Electroencephalograms (EEGs), Electrocardiograms (ECGs) and Electrooculograms (EOGs). Hence, they used these four categories to structure the content of their survey.

DNA sequencing – Deoxyribonucleic acid (DNA) sequencing is the determination procedure to reveal the order of nucleotides in DNA. Celesti et al. [38] review deep learning-based approaches to accelerate the process of DNA sequencing, given that huge amount of genomics data is emerging from next-generation sequencing (NGS) techniques. They provide a taxonomic analysis, by outlining the main deep learning-based NGS tools and software, and discuss

Table 5
Methods, pros, cons, challenges and future directions in medical deep learning in 2017.

Publication	Methods	Pros	Cons	Challenges	Future Directions
Medical image analysis I* Shen et al. [28]	CNN, DBM, DBN, SAE	-DL can learn features through labeled data itself -Can be used experts outside of the medical domain	-Overfitting due to limited training samples	-Image features learnt by deep learning are difficult to understand and interpret	-Build medical equivalent of ImageNET -Incorporate domain-specific knowledge in design/training -Develop a universal algorithm compatible with various imaging modalities and protocols
Healthcare Miotto et al. [29]	AE, CNN, RBM, RNN	-DL can model, represent and learn from heterogenous EHR	-Neural networks need improvement in interpretability, data integration, and security	-Low data volume -Data heterogeneity -Low interpretability -Domain complexity -Disease temporality	-Use of federated learning, explainable AI -Modeling temporality -Include expert knowledge into modeling -Preserve privacy -upscale and standardize EHR
Medical image analysis II* Litjens et al. [30]	AE, CNN, DBN, GAN, RBM, SAE, VAE	-End to end training (CNN) -Freely available pre-trained deep learning models	-Hyper-parameter tuning is empirical -Subjective medical image annotation is susceptible variability and uncertainty	-Medical image annotation is time consuming and expensive	-Task-specific pre-processing and data augmentation techniques -Incorporate prior knowledge of the specific domain into training -Radiological reports could be used to annotate medical images -Leverage non-expert annotation through crowd-sourcing -Unsupervised learning using unlabeled data -Interpretable DL
Stroke Management Feng et al. [31]	CNN, DNN	-Can apply automated featurization, image segmentation, multi-model prognostication, CAD	-Neural networks need improvement	-DL requires substantial programming skills -Data scarcity	-DL will increasingly become a personalized medicine tool for stroke specialists due to its speed, power and versatility
Analysis of molecular images in cancer Xue et al. [32]	AE, CNN, DNN, SAE	-Improved speed and performance in tumor segmentation, classification, and survival prediction	-CNN may overfit -CNNs have time consuming training, challenging with low data	-Insufficient and imbalanced datasets -Subjective model depth, architecture and hyperparameters -Abstract high-level features	-Self-supervised approaches can solve the annotation problem and make larger datasets usable -Explore model optimization and explainability -Establish larger-scale public datasets
Health Record Analysis Shickel et al. [33]	AE, CNN. MLP, RBM, RNN	-LSTM, RNNs, and variant can process sequential data	-Lack of transparency and interpretability	-Heterogenous data -Lack of reproducibility and universal benchmarks	-Include robust mechanisms to handle EHR irregularity -Focus NLP on the clinical notes -Unify the representation of various types of patients' data -Patient deidentification using DL -Increase interpretability
Microscopy image analysis Xing et al. [34]	CNN, FCN, RNN, SAE	-Unsupervised training (SAE) -Unfixed input size (FCN) -easily parallelized training (CNN)	-Obtaining large number of annotated microscopy images is expensive -NN requires a fixed input size	-Low interpretability -Processing high volumes of medical data require computational acceleration	-Develop DL methods for WSI analysis -Use a patch-based strategy to reduce computational expenses -Fusing different types of patients' data -Design task-specific DL architecture based on domain knowledge -Develop unsupervised or semi-supervised learning algorithms

Abbreviations: AE: auto-encoder, CAD: computer-assisted diagnosis, CNN: convolutional neural network, DBM: deep Boltzmann machine, DBN: deep belief network, DL: deep learning, DNN: deep neural network, EHR: electronic health record, FCN: fully convolutional network, GAN: generative adversarial network, MLP: multilayer perceptron, NLP: natural language processing, LSTM: long short-term memory, RBM: restricted Boltzmann machine, RNN: recurrent neural network, SAE: stacked auto-encoder, VAE: variational auto-encoder, WSI: whole slide imaging. *Also discussed in [74].

remaining research questions with a special focus on cloud computing.

Radiotherapy – Radiotherapy (or radiation therapy) utilizes ionizing radiation to control or kill malignant cancer cells. Therefore, treatment planning and delivery is complex and may be facilitated and partially automated by artificial intelligence. In their review, Meyer et al. [39] start explaining the fundamentals of deep learning-based techniques by relating them to the wider machine learning field. They give an overview of main network architectures, with special attention to convolutional neural networks. Afterwards, they analyze and summarize deep learning-based works for radiotherapy applications by classifying them into seven unique categories that are related to the workflow of the patient.

Ophthalmology – The diagnosis and treatment of eye disorders in medicine is called ophthalmology. In their review, Grewal et al. [40] explore deep learning as a new technology for ophthalmology with various possible applications. They explore deep learning-based methods that have been utilized in various diagnostic modalities, such as digital photographs, visual fields, and optical coherence tomography. They identify applications in the evaluation of numerous diseases, like cataracts, age-related macular degeneration, glaucoma, and diabetic retinopathy.

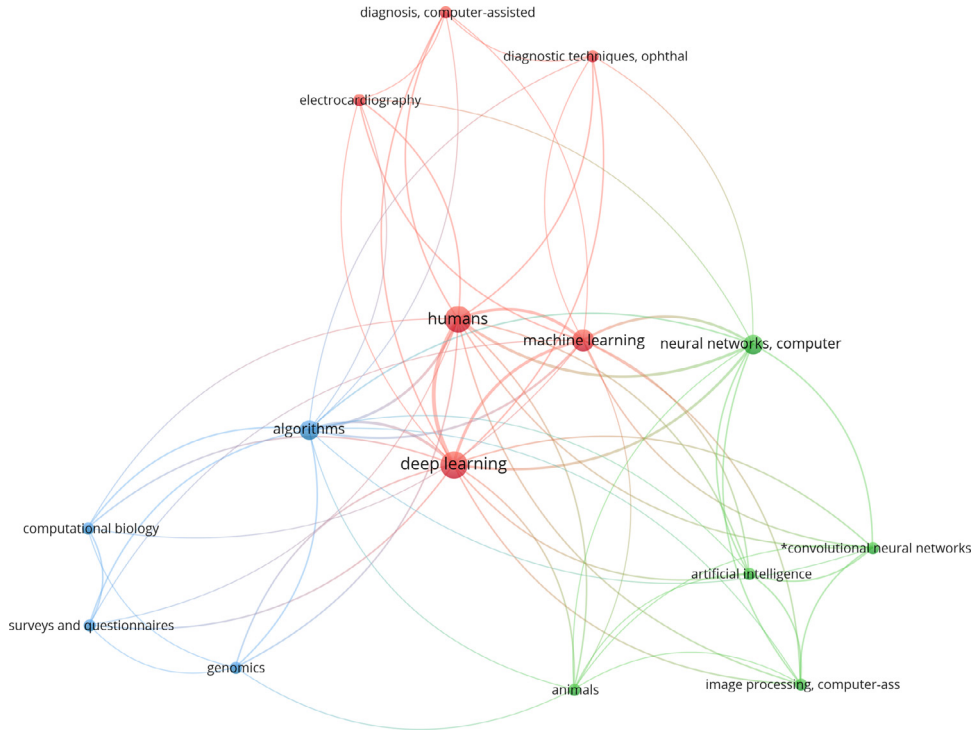

Fig. 4. Network visualization for the review articles supplied keywords in 2018 performed with VOSviewer.

Electronic health records – Electronic health records (EHR) summarize the data that is routinely collected from millions of patients across diverse healthcare centers, including information about patient demographics, diagnoses, medication prescriptions, clinical notes, laboratory test results, and medical images. Xiao et al. [41] performed a systematic analysis of deep learning-based models for exploring such EHR data by outlining them in regards to the kind of analytics task they perform and the kind of deep learning-based model architecture they use. They also depict the specific challenges resulting from such health data and tasks, and discuss potential solutions, as well as strategies for an evaluation in this field.

Bioinformatics – Bioinformatics is an interdisciplinary field developing approaches and software tools for the understanding of biological data with a strong focus on large and complex datasets. Lan et al. [42] survey research works combining deep learning-based methods with data mining, aiming to explore particular knowledge of the bioinformatics domain. The survey work gives a summary of several conventional algorithms in the data mining field that have been utilized for different tasks, like pre-processing, clustering and classification, but also of optimized neural network-based architectures and deep learning-based approaches. Finally, they outline the advantages and disadvantages in practical applications and discuss and compare them in terms of their industrial usage.

Personalized medicine – The aim of personalized medicine is to provide tailored patient-specific medical treatments via the identification of common features, like their inheritance, genetics and so on. Zhang et al. [43] provide a research outline concerning learning algorithms and methods, and their application, with an emphasis on deep learning-based approaches for personalized medicine. They explore three main application domains by giving insights into their pros and cons, namely disease characteristic identification, drug development, and a prediction of the therapeutic effect. They conclude that the analyzed learning algorithms and methods cannot be seen as a general solution for all kinds of medical problems.

1-D biosignals – Biosignals are electrical, thermal, mechanical or other signals measured over time, coming from the human body or other organic tissues, for example an ECG measures electrical activity originating from the heart muscle. Ganapathy et al. [44] survey deep learning approaches for 1-D biosignals in the field of computer-aided diagnosis. Further, they aim to establish a taxonomy to categorize the increasing number of applications in that area. The deep learning-based models were arranged according to the origin, type and dimension of the biosignal, the application goal, type and size of the ground truth data, type and schedule of network learning, and the overall model topology.

Omics – The emergence of big data has also involved the field of omics, including genomics, transcriptomics and proteomics. Zhang et al. [45] aim to give an entry-level overview, to understand the usage of deep learning approaches and methods for tackling problems and challenges in the omics domain. They outline and discuss various deep learning-based techniques that have fused deep learning with omics. Furthermore, they explore deep learning-based open-source frameworks with regard to their performances and features, but also highlight upcoming challenges and chances.

Sport-specific movement recognition – Sport-specific movement recognition can be utilized for the objective performance analysis of an (elite) athlete. In that regard, Cust et al. [46] explore the automated recognition and characterization of movements in sports, which can provide an alternative for an otherwise manual, time-consuming, limited performance analysis. The authors perform a systematical literature analysis on machine learning- and deep learning-based approaches for movement recognition in sports depending on input data from computer vision and inertial measurement units. They conclude that the experiment set-up, data preprocessing, and method development need to be considered and adjusted in accordance with the specific characteristics of the examined (sport) movements to achieve good results.

Diabetic retinopathy – Diabetes is on the rise worldwide and the most frequent microvascular complication is diabetic retinopathy, which can lead to visual impairment or even blindness. Nielsen et al. [47] performed a systematic review of deep learning techniques used for diabetic retinopathy screening. They explore works utilizing deep learning-based approaches for the classification of full-scale diabetic retinopathy, using retinal fundus images from diabetes patients. However, they only include works, which used a grading scale for diabetic retinopathy, a deep learning performance score, and have been compared to a reference standard from a human grader.

Image cytometry – Cytometry is the measurement of cell characteristics, like the cell size, cell count, cell morphology, cell cycle phase and DNA content. Gupta et al. [48] review how deep learning has been used to analyze microscopy image data of tissue samples and cells. They begin with an overview of neural networks and deep learning. They outline requirements for the input data, computational resources, and limitations and challenges in published works on deep learning in image cytometry, as well as identify methods that have not yet been used for cytometry data for potential future work.

Radiology – Radiology is the medical field of extracting useful information from images, like computed tomography (CT) or MRI, for diagnosis and treatment of humans and animals. In their review, Mazurowski et al. [49] give an introduction about the field of radiology and outline open research questions that could be tackled with deep learning techniques. They further provide an overview of basic deep learning concepts, such as convolutional neural networks. Next, they outline deep learning-based research contributions published within the radiology discipline. Thereby, they organize the reviewed works by the specific type of tasks they aim to support. They conclude their work by discussing remaining problems, but also highlight opportunities for using deep learning-based approaches within the practice of radiology.

2.2.1. Diving deeper: architectures, evaluations, pros, cons, challenges and future directions in 2018

Table 6 presents more details about the presented methods, pros, cons, evaluations and challenges and future directions for the reviews from the year 2018. Again, all the reported reviews share several important conclusions. Deep learning methods outperform machine learning methods over a wide variety of subjects, tasks, and datasets. All reviews predict an increase in (and increasing importance) of deep learning-assisted research and, at some future junction, practical applications. Most deep learning methods covered in the individual papers were CNNs and the review authors specifically cite CNNs as yielding impressive automatically extracted features and performances. The same general issues that were already reported in 2017, such as lack of interpretability and high-quality dataset availability, are reported again. Lastly, while generally considered promising, deep learning methods at this point have not been integrated into practical workflows.

2.3. Medical deep learning reviews in 2019

With the proposed search strategy, 21 surveys were identified in the area of medical deep learning in 2019. Fig. 5 shows a network visualization for the review articles supplied keywords in 2019 that reveals the keyword “deep learning” and its connections as the main cluster. Further main keyword clusters are “humans”, “machine learning”, and “artificial intelligence”. New clusters arise around the keywords “brain” and “brain-computer interfaces”, which shows that this organ has been heavily targeted by the research community in 2019. Also interesting is the cluster around “convolutional neural network”, which shows that CNN gained momentum in the medical domain by 2019. The proposed reviews from 2019 refer to 2279 contributions and have already been cited 408 times (status as of August 2020). They are ordered by *epub* date in 2019 (Table 4) and cover the following categories:

- Medical imaging;
- Brain cancer classification;
- Electroencephalogram;
- Pulmonary nodule detection;
- Neuro-oncology;
- Diabetic retinopathy;
- Cardiac arrhythmia;
- Protein structure;
- Electroencephalography;
- Neurology;
- Cancer diagnosis;
- Ultrasound;
- Radiation oncology;
- Drug-drug interaction;
- Urology;
- Sleep apnea;
- Ophthalmic diagnosis;
- Alzheimer’s disease;
- Pulmonary nodule detection;
- Liver masses;
- Pulmonary medical imaging.

Medical imaging – Medical imaging covers the field of producing visual representations of the internal body, for example using computed tomography, magnetic resonance imaging or ultrasound, just to name a few. Biswas et al. [50] explore various types of deep learning systems available, with a focus on current deep learning-based applications in medical imaging. They also outline the transition of technology from machine learning to deep learning and provide a complexity analysis and potential advantages for developers and users.

Brain cancer classification – In general, brain tumors are classified into several types, depending on whether they are, for example, benign or malignant, which helps to choose an optimal treatment for the patient. Tandel et al. [51] review machine learning and deep learning-based methods in the field of brain cancer, with a focus on pathophysiology. They include a review of imaging modalities and automatic, computer assisted methods for the characterization of brain cancer. Moreover, they outline the analysis of connections between cancer in the brain and additional brain disorders, such as Alzheimer’s disease, Wilson’s disease, Parkinson’s disease, stroke, leukoaraiosis, and further neurological disorders.

Electroencephalogram – In the field of neuroscience, EEG analysis is an important technique with applications not only in neuroscience, but also neural engineering, like brain-computer interfaces (BCIs). Craik et al. [52] perform a systematic review on deep learning applications for EEG classification, addressing several questions, including specifiying specific EEG tasks. They analyze the studies based on several categories, like preprocessing algorithms for EEG, the kind of input, and the type of deep neural network architecture. The deep learning tasks were divided into five groups, namely the mental workload, emotion recognition, seizure detection, motor imagery, event related potential detection, and sleep scoring. For every kind of task, they outline the specific formulation of the input, classifier recommendations, and other major important characteristics.

Pulmonary nodule detection – Pehrson et al. [53] systematically reviewed the deep learning or machine learning-based methods used for the automatic detection of pulmonary nodules using a common dataset, the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database. They divide the works into two subcategories based on their overall architecture.

Table 6
Methods, pros, cons, challenges and future directions in medical deep learning in 2018.

Publication	Methods	Pros	Cons	Challenges	Future directions
Toxicity of chemicals Tang et al. [35]	DNN	-DNNs outperform ML in prediction of epoxidation, quinone formation, metabolite reactivity, classification of toxicity effects, and chemical-target interaction prediction	-Overfitting in DNNs -DNN underperformance with small quantities of training data	-Imbalanced, inhomogeneous, small datasets -Necessity of long training times and large computational resources	-Creation and curation of larger, public datasets by combining datasets from published works, patents and the web
Pulmonary nodule diagnosis Yang et al. [36]	AE, CNN, DBN, MTANN, SDAE	-Inclusion of CAD, feature extraction and benign-malignant classification -CNNs outperform SVMs and handcrafted rule-based algorithms	-Large amounts of data are required for successful training	-To facilitate DL, large datasets must be set up using time-consuming and unreliable manual labeling	-DL for decision support in pulmonary nodule diagnosis and classification -Alleviate the burden of dataset labeling with reinforcement learning -Create public datasets similar to ImageNet -Multi-scale patches during training to bridge data gap
Physiological signals Faust et al. [37]	AE, CNN, DBN, DNN, KNN, LSTM, RBM, RNN, SDAE, SVM	-Eliminates tedious and error-prone manual feature selection -Successful applications include state predictions, classifications and signal decoding.	-Time consuming -Model architecture and hyperparameters decided without statistical evaluation -Failure to capture information in a generalizable way for chaotic signals	-Long training times -Need for large training sets	-Testing DL applications in practical settings
DNA sequencing Celesti et al. [38]	AE, CNN, DNN, HMM, MLFF, RNN	-Integrated into software for gene expression analysis, genome analysis, SNP research, and early cancer detection -Computational efficiency and best performance/generalization	-Not discussed	-Most existing NGS library preparation devices, sequencing instruments, and software tools have not been designed to work in a clinical networked environment	-DL for comparative genomics, forensic biology, biological systematic field, virology) -Cloud computing services will provide scalability and data sharing possibilities
Radiotherapy* Meyer et al. [39]	AE, CNN, DNN, RNN	-Availability of large amount of training data -Increasing power of GPUs	-DL theories are empirically and experimentally obtained -Small noise, imperceptible to humans, could alter the output completely	-Building coherent, large and balanced medical datasets that represent real-world scenarios -Difficulty of interpretation	-Not discussed
Ophthalmology Grewal et al. [40]	CNN, others unnamed	-DL has superior performance compared to older automated methods -Successful application for early diagnosis of age-related macular degeneration, diabetic retinopathy, glaucoma	-Difficulty conveying quantitative results (such as disease severity) -Overfitting on uncorrelated features, noise, or dataset-inherent biases	-Overinterpreting results from neural networks -Variability in dataset labels, and medical definitions	-Retinal photography with smartphones and DL deep learning could enable self-ophthalmology and diagnoses -Integrate DL in the ophthalmologic routine
Electronic health records Xiao et al. [41]	AE, CNN, GAN, GRU, LSTM, RNN, UE	-DL: better performance and less manual feature engineering required -Availability of large and complex datasets in healthcare for training -Successfully applied to clinical event prediction, disease classification, phenotyping, text labeling, generating continuous medical time series	-Lack of interpretability	-Temporality and irregularity of EHR data with lack of labels and multi-modality -Lack of generalization	-Interpretable and transparent model creation and data curation
Bioinformatics* Lan et al. [42]	CNN, DBN, decision tree, DNN clustering, NB, KNN, RNN, SAE, SVM	-DL can learn knowledge from massive amount of data automatically	-DL requires large datasets for training -Dependent on high-end hardware -Lack of interpretability	-Data imbalance is prevalent in the medical domain	-Aggregate different ML algorithms -Fuse data from different modalities -Develop semi-supervised and reinforcement learning algorithms
Personalized medicine Zhang et al. [43]	ANN, Bayesian networks, CNN, DBN, DNN, linear regression, MLP, RF, SDAE, SVM	-More modern DNNs and CNNs outperformed older algorithms -Scale more efficiently with increasing dataset complexity -Feature recognition and structural association in structured data -Successfully applied for drug development, disease characteristics and therapeutic effects	-Have not been applied to large scale datasets -Human intervention is required to extract new knowledge and for safe action	-Dataset limited availability, uncertainty, idiosyncrasy, size -Lack of reproducibility overfitting, computational complexity -Data privacy, lack of clinical approval, intellectual property rights, genetic correlation validation	-Upgrade clinical data and integration of already developed algorithms -Develop more reliable automated feature selection -Field growth

(*continued on next page*)

Table 6 (*continued*)

Publication	Methods	Pros	Cons	Challenges	Future directions
1-D biosignals Ganapathy et al. [44]	AE, ANN, CNN, DBN, DNN, RBM, RNN	-Non-linearity and complexity handled well -Good performance even with multi-modal or complex data -Successfully applied to enhancement, detection, clustering, diagnostics, and prediction.	-Weaknesses not explicitly covered, only the inherent challenges	-Small and complex datasets, device specificity, noise -Real-time requirements for clinical applications -Missing ground truths	-Increase standardization of network topology and parameters
Omics Zhang et al. [45]	CNN, DBN, DNN, GRU, LSTM, MLP, RBM, RNN, SAE	-Successfully applied to DNA, RNA, protein structure analysis, gene expression regulation analysis, disease prediction, protein function analysis -CNNs can analyze spatial information in images -RNNs can analyze correlated features and time-series -DNNs are highly adaptable to almost all types of data	-Older RNNs are unstable during training -Data cleaning is time-consuming and labor-intensive -More training data, computations resources, and higher data quality required -Lack of interpretability	-Model selection and parameter tuning	-Increasing relevance of reinforcement learning, incremental learning, and transfer learning -Mitigation techniques for the disadvantages of DL methods will continually be developed
Sport-specific movement recognition Cust et al. [46]	CNN, DTW, KNN, LSTM, MLP, HMM, NB, RF, SVM	-DL outperforms other ML methods in performance and computational efficiency -Does not rely on heuristic features	-Not discussed	-Lack of uniformity in data acquisition	-Fusion of IMU and vision data in models
Diabetic Retinopathy Nielsen et al. [47]	CNN, DNN	-Reduced manpower due to automation, cost of screening, and issues relating to interrater reliability	-Lack of trust due to "black box" nature	-Risk of bias towards favorable results due to exclusion of difficult images from datasets -Lack of interpretability	-Overcome challenges with prediction uncertainty, quality control and lack of interpretability
Image cytometry Gupta et al. [48]	AE, CNN, DNN, GAN, MLP, RNN	-Features are generated independently and automatically -Use of "transfer learning" -Successful application areas covered all modalities, tasks and scales	-Require large amounts of annotated data -Lack of interpretability -Overfitting and underfitting	-Requires computational resources and programming expertise -Class imbalances can impede the generalization ability -Lack of interpretability	-Combine hand-crafted features and neural network analysis for strong, grounded results
Radiology* Mazurowski et al. [49]	ANN, CNN	-Effective in medical image classification, segmentation, detection, reconstruction and registration	-DL only outperformed human experts in a minority of radiological tasks -Introducing DL into clinical practice will cause legal and ethical issues	-Datasets are smaller and often imbalanced, leading to suboptimal training -Proper clinical validation is often overlooked	-Optimally incorporate DL in existing radiology workflow

Abbreviations: AE: auto-encoder, ANN: artificial neural network, CAD: computer-assisted diagnosis, CNN: convolutional neural network, DBN: deep belief network, DL: deep learning, DNN: deep neural network, GAN: generational adversarial networks, GPU: graphic processing unit, GRU: gated recurrent units, HMM: hidden Markov model, IMU: inertial measurement unit, KNN: K-nearest neighbors, LSTM: long short-term memory, ML: machine learning, MLFF: multi-layer feed forward, MLP: multi-layer perceptrons, MTANN: massive training artificial neural network, NB: Naïve Bayes, NGS: next-generation sequencing, RBM: restricted Boltzmann machine, RF: random forest, RNN: recurrent neural network, SDAE: stacked denoising auto-encoder, SNP: single nucleotide polymorphism, SVM: support vector machine,.
UE: unsupervised embedding, *Also discussed in [74].

They conclude that machine learning and deep learning methods can be used for the detection of lung nodules, even with a high level of sensitivity, specificity and accuracy, however, they also conclude that there is no general technique to evaluate the performance of machine learning methods and algorithms.

Neuro-oncology – Gliomas represent 80% of all primary malignant brain tumors. Shaver et al.'s [54] survey provides an overview of the recent deep learning-based approaches and applications utilized for glioma detection and outcome prediction. They focus on the pre-operative and post-operative segmentation of tumors, genetic tissue characterization, and further prognostication. They show and conclude that deep learning-based approaches and applications are promising research directions for the segmentation and characterization of gliomas, their grading, and for giving a survival prediction.

Diabetic retinopathy – Another survey about diabetic retinopathy was published by Asiri et al. [55]. They focus on deep learning-based computer-aided diagnosis (CAD) systems, which they structure into various stages such as lesion segmentation, lesion detection, and lesion classification of fundus images. Furthermore, they discuss pros and cons of published deep learning-based methods to accomplish these tasks.

Cardiac arrhythmia – Cardiac arrhythmias are most commonly detected by an ECG, mainly because of its low cost and convenient usage. For these reasons, every day, ECG data is acquired in large amounts in hospitals and homes, which, on the downside, prevents a detailed manual data inspection. Parvaneh et al. [56] perform a review of recent advancements on cardiac arrhythmia detection using deep learning. They outline existing works according to five different aspects, namely the used dataset, the input data type, the kind of application, the applied architecture model, and finally, the evaluation of performance. They conclude by presenting the shortcomings of the surveyed studies and discuss possible future upcoming research directions.

Protein structure – The three-dimensional form of local segments of proteins is called protein secondary structure. Wardah et al. [57] wrote a review on predicting the secondary structures of proteins with deep learning-based approaches such as neural networks. They start with a background section about the secondary structure of a protein and introduce the basics of artificial neural

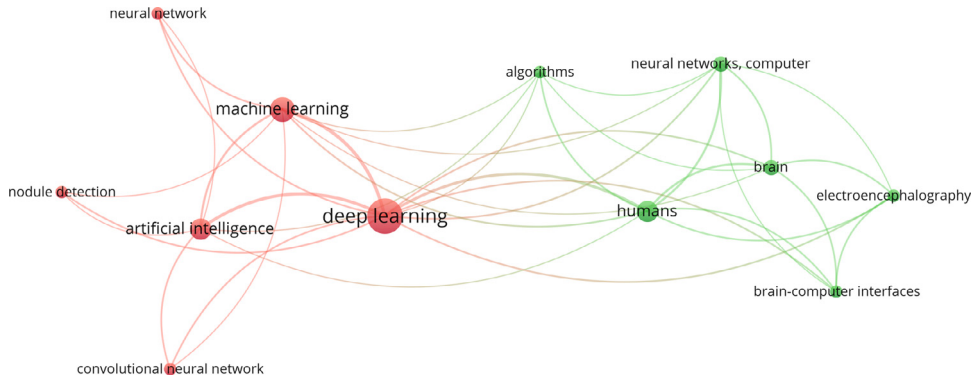

Fig. 5. Network visualization for the review articles supplied keywords in 2019 performed with VOSviewer.

networks. The authors conclude that there are several challenges left for the *in silico* predictions of secondary protein structures.

Electroencephalography – As stated beforehand in the review about EEG classification, EEG analysis is an important yet difficult task, which requires several years of training because of its complexity. Roy et al. [58] performed a systematic survey of the analysis of electroencephalography with deep learning methods, covering various applications domains, like sleep, epilepsy, cognitive and affective monitoring, and brain-computer interfacing. In addition, they collected information pertaining to the data, such as the pre-processing methodology, the selection of the deep learning design, the results, and the experiments' reproducibility.

Neurology – The medical branch related to nervous system disorders (central and peripheral) is named neurology. Neurology covers the diagnosis and treatment of such disorders. Valliani et al. [59] review various neurology domains where deep learning algorithms have already been applied, like Alzheimer's disease diagnosis and early acute neurologic event detection. They also survey the segmentation of medical images for a quantitative evaluation of the neuroanatomy and vasculature structure, connectome mapping for Alzheimer's diagnosis, autism spectrum disorder (ASD), and attention deficit hyperactivity disorder (ADHD), as well as explore the granular genetic signatures and the signals of microscopic electroencephalograms.

Cancer diagnosis – A range of diseases, involving an abnormal growth of cells, which can also spread and invade other parts of the body, is called cancer. Munir et al. [60] give a bibliographic analysis on cancer diagnosis with deep learning-based approaches, starting with a background description of the cancer diagnosis domain. They cover the individual steps for cancer diagnosis, but also classification methods, like the asymmetry, border, color and diameter (ABCD) method, the Menzies method, the seven-point detection method, and pattern analysis. For each reviewed deep learning technique, they link to Python code. They also compile the applied deep learning models for different cancer types. Specifically, they discuss brain cancer, breast cancer, skin cancer and lung cancer.

Ultrasound – Ultrasound (US) is commonly used in the clinical routine due to it is nonionizing, low-cost, and portable characteristics, coupled with the ability of providing real-time images. Akkus et al. [61] present a review on deep learning-based applications in the ultrasound domain with the aim to improve the clinical workflow, including improving the acquisition of the US images, real-time evaluation image quality, objective detection and disease diagnosis, and in general, an overall optimized clinical workflow during ultrasound examinations. They also give a specific forecast of upcoming research trends and directions for deep learning-based methods that can facilitate an US diagnosis, but also reduce costs in health care, and provide an optimized clinical US workflow.

Radiation oncology – A physician or doctor who is specialized in the treatment of cancer using ionizing radiation, like radionuclides or megavoltage X-rays, is called a radiation oncologist. In that context, Boldrini et al. [62] perform a literature review in PubMed/Medline with a search strategy including the search terms "radiotherapy" and "deep learning". They identify recent publications on deep learning in radiation oncology, which they present with a focus on clinically oriented readers. The review shows how deep learning can support clinicians during their daily work, such as by reducing segmentation times, or predicting treatment outcomes and toxicities. However, they conclude that these techniques have yet to be employed in the clinical routine, and it remains to be seen how well they translate into practice.

Drug-drug interaction – Drug-drug interactions (DDIs) can cause adverse drug effects that have the potential to threaten public health and patient safety. Hence, these interactions are crucial for drug research and pharmacovigilance. Zhang et al. [63] review the state-of-the-art deep learning-based methods used for DDI exploration. They briefly outline every deep learning method from their surveyed studies and systematically evaluate their efficiency, strengths and weaknesses. They conclude their work by providing a discussion and giving an outlook on several future research challenges for the extraction of DDIs with deep learning-based approaches.

Urology – The medical branch of urology is focused on surgical and medical urinary-tract system diseases, including the urethra, urinary bladder, ureters, adrenal glands and kidneys. The urology branch also focuses on the reproductive organs of males, including the prostate, testes, penis, epididymis, seminal vesicles and vas deferens. Suarez-Ibarrola et al. [64] review recent and upcoming machine learning- and deep learning-based applications in the urology domain, with a focus on renal cell carcinomas, urolithiasis, prostate and bladder cancer. This covers, for example, the prediction of endourologic surgical outcomes in urolithiasis, the automatic distinction between malignant and benign small renal masses, the analysis of texture features and radiomics for the differentiation between low-grade and high-grade tumors in bladder cancer, MRI-based computer-aided diagnosis, biochemical recurrence prognosis, and prognosis algorithms for the Gleason score for prostate cancer.

Sleep apnea – Sleep apnea is a sleep disorders characterized by repeated stopping and starting of breathing. Sleep apnea can be scored with polysomnography, which is unfortunately expensive, inaccessible, uncomfortable and requires an expert technician. Mostafa et al. [65] preform a systematic review on the published deep learning-based research contributions used for detecting sleep apnea. They focus on exploring research subjects including the implementations of neural networks, a possible need for pre-processing or manual feature extraction, and finally, explore

Table 7
Methods, pros, cons, challenges and future directions in medical deep learning in 2019.

Publication	Methods	Pros	Cons	Challenges	Future Directions
Medical Imaging Biswas et al. [50]	AE, (fully) CNN, DBN, DRN, FCN, SVM	-DBM has easy inference -Automated feature extraction -Learning of complicated and composite relationships in data -DL methods surpass in robustness and performance	-Unknown generalization capabilities (DBN) -Vanishing gradient problems during training (AE)	-Improvement needed before techniques could be integrated into clinical workflows -Only trained on small datasets	-Widespread use in research and clinical routine -Develop real-time applications
Brain cancer classification Tandel et al. [51]	ANN, CNN, EM, KNN, NB, RF, SVM	-Automatically produce features that are stable to deformation and translation invariant -DL outperforms other ML methods	-Computationally more expensive	-Not discussed	-Provide the fast, non-invasive diagnosis tool that the field needs
Electroencephalogram Craik et al. [52]	AE, CNN, DBN, LSTM, MLP, RBM, RNN, SAE, SVM	-Successfully applied to motor imagery, seizure detection, mental workload, sleep stage scoring, event related potential, and emotion recognition	-Not discussed	-Formulation of the input data (PSD, wavelet decomposition, etc.)	-Combine convolutions and recurrent or RBM architectures -Use de-noised EEG data
Neuro-oncology Shaver et al. [54]	ANN, CNN, CRNN, LSTM, SVM	-Do not require human-constructed features -CNN architectures provide high accuracies on segmentation, characterization, grading and survival prediction tasks	-Requires large quantities of annotated data, necessitating medical expert knowledge and significant amounts of time -Overfitting	-Lack of large amounts of annotated data	-Undisruptive integration into workflows -Work with regulatory bodies who currently restrict the use of ML/DL in clinical practice
Diabetic retinopathy Asiri et al. [55]	AE, CNN, DBN, RNN	-Automatic discovery of relevant features -Ability to train and deliver solutions in an end-to-end manner -Successfully applied to vessel and optic disk segmentation, lesion detection and classification, diabetic retinopathy diagnosis	-Require large amounts of labeled data -Tendency to overfit -Convergence of DL methods not always guaranteed -Lack of interpretability -Class imbalance of datasets	-Lack of large-scale annotated uniform training data -Generalization of DL methods	-More standardization in data, labels, and test metrics -Research GANs
Cardiac arrhythmia Parvaneh et al. [56]	AE, CNN, DBN, LSTM, RNN	-Unsupervised information capture and feature generation	-Highest scoring ML outperformed best DL -Overfitting	-Lack of interpretability -Large datasets needed	-Research interpretability -Identify optimal dataset sizes for training and testing
Protein structure Wardah et al. [57]	ANN, CNN, GRU, HMM, RNN	-Automatic protein structure prediction -Reduced time and costs compared to traditional in vitro analysis	-Not discussed	-Need in vitro techniques to determine hard truths, limiting datasets -Lack of comparability	-Automated prediction methods will drive the benchmark in the field closer to the theoretical accuracy boundary (approx. 88%)
Electrocephalography Roy et al. [58]	AE, CNN, DBN, GAN, MLP, RBM, RNN, SDAE	-Avoids time-consuming traditional feature engineering and provides end-to-end solutions -Can flexibly work with either small or large amounts of data -Can generalize to other tasks or datasets -Successfully applied to tasks including brain–computer interfacing, sleep staging, epilepsy, cognitive and affective monitoring	-Lack of reproducibility and interpretability	-Lack of labeled data -Dataset augmentations and hyperparameter searches are difficult to identify	-Efforts in reproducibility -Exploratory research into data quantity vs performance
Neurology Valliani et al. [59]	AE, CNN, DNN, GAN, GRU, LSTM, NB, RNN, SVM	-No manual feature crafting -Performance gains with larger datasets -Successfully applied for medical image classification, segmentation, functional connectivity, classification of brain disorders and risk prognosis	-Require large amounts of data to learn -High quality labels are time-consuming to create -Overfitting -Lack of interpretability	-Medical data suffers from heterogeneity and complexity -Data privacy, accessibility and ethical concerns over potential biases	-Research into generalizability and interpretability
Cancer diagnosis Munir et al. [60]	AE, AFINN, (fully) CNN, DBN, GAN, LSTM, RBM, RNN	-Learn features from raw images instead of requiring manually constructed features -Successfully applied to cancer diagnosis on multiple image modalities	-Require large datasets, generally with labels, a major time/cost investment	-Lack of available datasets -Datasets suffer from a strong disparity between positive and negative samples	-Not discussed
Ultrasound Akkus et al. [61]	AE, (fully) CNN, RBM, RNN, SDAE, SVM	-DL outperformed ML in generalizability -Successfully applied to detection, classification, segmentation, and diagnosis of lesions and nodules	-Lack of interpretability and explainability	-Dataset quality and performance vary in acquisition and interpretability -Size and quantity of public datasets are limited	-Clinical workflow and cost can be reduced -Include 3D, multiview cine clips, or spatiotemporal data into AI models

(*continued on next page*)

Table 7 (*continued*)

Publication	Methods	Pros	Cons	Challenges	Future Directions
Radiation Oncology Boldrini et al. [62]	ANN, (fully) CNN, DNN, GAN, SVM	-Can analyze unstructured data and extract non-linear features without human supervision -Capable of dimensional reduction -Successfully applied to segmentation, outcome, response, and survival predictions	-Loss of functions are non-convex and no algorithm can guarantee to find an optimal solution -Overfitting	-Need for expert knowledge in oncology and DL for dataset curation and training	-Need for bigger standardized datasets
Drug-drug interaction Zhang et al. [63]	CNN, GRU, LSTM, RNN, recursive neural network	-No need for manual feature engineering -CNNs can generate translation-invariant descriptions from data -RNNs can selectively hold relevant information in memory and analyze arbitrary length text inputs	-Tendency to be unstable during training -Lack of interpretability	-Unstructured data and class imbalances	-Semi-/self-supervised learning, joint learning models, N-ary relation extraction, feature enrichment, interpretable modeling
Urology Suarez-Ibarrola et al. [64]	ANN, CNN, SVM	-Details not discussed	-In some cases, ML/DL were favorable to human raters, but traditional statistical methods outperform them, particularly in the field of urolithiasis	-Equipment variants and non-standardized data collection -Generalization -Heterogeneity of employed models and datasets	-Create large-scale public datasets -Keep downscaling in mind to employ DL methods in real-time or on mobile devices
Sleep Apnea Mostafa et al. [65]	CNN. DBN, GRU, LSTM, MLP, RNN, SSAE	-Increased performance of DL vs ML methods	-Details not discussed	-Imbalanced heterogenous datasets -Hyperparameter search	-Not discussed
Ophthalmic diagnosis Sengupta et al. [66]	(fully) CNN, FNN, MBNN, RF, SSAE, SVM	-DL outperforms for lesion and vessel segmentation, acute macular degeneration, glaucoma and diabetic retinopathy classification	-Requires large amounts of annotated data for training -Can suffer from domain shift between training and test sets -Generalizability	-Class imbalance -Data acquisition and performance indicators are heterogeneous across reported papers	-Research generative models to augment existing datasets or balance classes -Domain adaptation
Alzheimer's disease Ebrahimighahnavieh et al. [67]	AE, CNN, DBN, DNN, DPN, HMM, DBM, RBM, SVM	-Suited for modeling non-linear relationships -Robust against translation and transformations of target features -Capable of automated feature generation	-Require large amounts of data for training -Loss of generalization capability -Overfitting, computational cost and robustness	-Unpublished code bases -Dataset imbalances and lack of data -ROI-based methods require extensive domain expert knowledge	-Public benchmarking platform for fair comparisons of models -Explainable AI -Generation methodology
Pulmonary nodule detection Li et al. [68]	(MT)ANN, CNN, SDAE	-MTANNs and SDAEs can learn with fewer training examples than CNNs and generate new data easily -Successfully applied to detection and classification of pulmonary nodules	-Longer training times and greater dataset requirements -Small datasets in medicine -Overfitting	-Heterogeneity of results	-Research into consistent, standardized integration of DL into clinical workflow
Liver masses Azer [69]	(fully) CNN, GAN	-Successfully applied to detection, classification, and segmentation of liver masses	-Details not discussed	-Heterogeneity of results	-Standardize reporting, report multiple performance metrics, practically apply, reproduce -Collaborative data acquisition -Case control studies to compare DL methods with human raters
Pulmonary medical imaging Ma et al. [70]	ANN, (fully) CNN, DPN, neural hypernetwork	-Self-learning and generalization -Can extract information both from simple and complex data structures	-High computational and dataset size requirements -Lack of interpretability	-Class imbalances in datasets -Varying image quality	-Make use of unlabeled medical data to ease the annotation bottleneck -Develop more interpretable DL models

Abbreviations: AE: auto-encoder, AFINN: adaptive fuzzy inference neural network, ANN: artificial neural network, CNN: convolutional neural network, CRNN: convolutional recurrent neural network, DBN: deep belief network, DBM: deep Boltzmann machine, DL: deep learning, DNN: deep neural network, DPN: dual path network, DRN: deep residual network, EM: expectation maximization, FCN: fully connected network, FNN: feed-forward neural network, GAN: generational adversarial networks, GRU: gated recurrent units, HMM: hidden Markov model, KNN: K-nearest neighbors, LSTM: long short-term memory, MBNN: Multi-branch neural network, ML: machine learning, MLP: multi-layer perceptrons, MTANN: massive training artificial neural network, NB: Naïve Bayes, RBM: restricted Boltzmann machine, RF: random forest, RNN: recurrent neural network, SAE: stacked auto-encoder, SSAE: Stacked sparse auto-encoder, SVM: support vector machine.

the reported applications in terms of implementation and performance. The applied sensors, signals, databases and implementation difficulties have also been taken into consideration for an automatic, deep learning-based scoring process.

Ophthalmic diagnosis – Sengupta et al. [66] provide another review on ophthalmology, focusing on ophthalmic diagnosis using deep learning approaches based on fundus images (the back surface of the eye). They discuss recent deep learning approaches for diabetic retinopathy, glaucoma and age-related macular degeneration, and describe numerous datasets consisting of retinal images, which can be processed for deep learning-based ophthalmic tasks. Areas of applications from their surveyed works include segmen-

tation of the optic cup, optic disk, and blood vessels, as well as lesion detection.

Alzheimer's disease – In developed countries, Alzheimer's Disease (AD) is one of the leading causes of death. AD is a chronic neurodegenerative disease that often starts slowly, but progressively worsens in the long-term. In this regard, Ebrahimighahnavieh et al. [67] systematically reviewed deep learning-based methods for an automatic AD detection from neuroimaging. They focus on the extraction of effective features and biomarkers, like genetic data, personal information, and scans of the brain, as well as required pre-processing steps and tips for handling neuroimaging data that comes from single- or multi-modality investigations. Moreover, they compare the performance of the deep learning models in AD detection and discuss remaining challenges, including the applied training strategies and datasets that can be accessed.

Pulmonary nodule detection – Another systematic review on deep learning-based methods in pulmonary nodule detection was published by Li et al. [68]. They focus on the detection and classification of nodules using CT scans not from the LIDC-IDRI database. They found that three types of deep learning architectures are commonly used, namely convolutional neural networks, deep stacked denoising autoencoder extreme learning machine (SDAE-ELM) methods and massive training artificial neural networks (MTANN). They conclude that high accuracy, specificity and sensitivity scores can be obtained with deep learning-based approaches in nodule classification and detection using CT scans not from the LIDC-IDRI cases.

Liver masses – A liver mass is a lesion in the liver that can be caused by an abnormal cell growth, a cyst, hormonal changes, or an immune reaction, but is not necessarily cancer. Azer [69] performs a systematic analysis on deep learning-based approaches, specifically convolutional neural networks (CNNs), for the detection of liver masses as well as hepatocellular carcinomas (HCCs). PubMed, the Web of Science, EMBASE and further research books were searched systematically, thereby identifying works analyzing cellular images, pathological anatomy images, and radiological images of liver masses or HCCs. The level of accuracy and CNN performance in cancer detection were presented with a focus on analyzing the kinds of liver masses and cancers and determining the image types which proved optimal for the precise detection of cancer.

Pulmonary medical imaging – Ma et al. [70] present an analysis on deep learning-based approaches for pulmonary medical imaging. Topics include classification, detection, and segmentation tasks in regard to pulmonary medical images, but also benchmarks and datasets. They provide an outline of the reviewed approaches, which have been implemented for different diseases of the lung, such as pneumonia, pulmonary embolisms, pulmonary nodules, and interstitial lung disease (ILD). Finally, they discuss the future challenges and potential directions in the area of medical imaging with deep learning techniques.

2.3.1. Diving deeper: architectures, evaluations, pros, cons, challenges and future directions in 2019

Table 7 presents more details about the presented methods, pros, cons, evaluations and challenges and future directions for the reviews from the year 2019. Interestingly, while most reviews cite largely the same advantages and disadvantages for deep learning, authors occasionally disagree on whether specific aspects of neural networks pose advantages, disadvantages or challenges, particularly concerning data availability. Some studies have had success training with very small datasets, while others did not, suggesting that not all the nuances of data pre-processing, augmentation, and training processes are fully understood yet. Many reviews report that individual papers could not be fairly compared in terms of performance due to the heterogeneity of methods and key performance indicators used, as well as due to the manifold differences in both datasets and data acquisition between reported papers. There appears to be a significant research gap in terms of standardization for these issues. Sometimes simpler statistical methods or traditional machine learning outperform deep learning and occasionally deep learning is reported to work better when shallower architectures are used, but typically deep learning methods handily outperform any competitors except human raters with years of experience. CNN architectures are typically used/reported the most often in the various review papers and many authors specifically report that CNNs appear to dominate the field both in terms of performance and prevalence. Lastly, deep learning methods are not deployed in clinical practice despite regularly achieving state-of-the-art results. Authors typically cite ethical concerns due to lack of interpretability, potential lack of generalizability, and unknown (or unknowable) biases as the reason. Thus, practical applications and real-world performance testing of newly developed deep learning methods, as well as deeper investigations into Explainable AI, constitute significant research gaps.

3. Conclusion

In this work, reviews and surveys on medical deep learning are presented in a systematic meta-review contribution. A systematic search has been performed in the common medical search engine PubMed, which resulted in over 40 review or survey publications published during the last three years. In addition to a brief summary of each survey, the references and citations of these reviews are presented (status as of August 2020).

Before 2017, no medical deep learning review article has been indexed under PubMed according to the proposed search strategy. This is easily explainable, because even though these kind of approaches had already been suggested and applied at the end of the last century [71,72], deep learning-based approaches only started to gain massive popularity after the convolutional neural network architecture AlexNet [73] won the ImageNet challenge in 2012. From that moment on, deep learning and convolutional neural networks have received inexorably increasing attention in various communities, including medical image analysis. However, it took some time to have enough published works for the first review or survey articles. In addition, there is also a massive number of review and survey articles in other, general disciplines. To give a rough impression of these, we performed an additional non-systematic search, which is, however, far from complete and the results are only presented in a systematic listing, because these works would go far beyond the scope of this contribution. Nonetheless, they may be an inspiration for interested readers and we arranged them in three categories (more details can be found in [74]):

1. Computer vision

   Object detection [75–77]
   Image segmentation [78,79]
   Face recognition [80–82]
   Action/motion recognition [83,84]
   Biometric recognition [85,86]
   Image super-resolution [87]
   Image captioning [88]
   Data augmentation [89]
   Generative adversarial networks [90]

2. Language processing

   General language processing [91]
   Language generation and conversation [92–95]

Named entity recognition [96,97]
Sentiment analysis [98,99]
Text summarization [100]
Answer selection [101]
Word embedding [102,103]
Financial forecasting [104]

3. Further works

Big data [105–107]
Reinforcement learning [108–110]
Mobile and wireless networking [111]
Mobile multimedia [112]
Multimodal learning [113]
Remote sensing [114]
Graphs [115]
Anomaly detection [116]
Recommender systems [117]
Agriculture [118]
Multiple areas [119–121]

4. Discussion

Typically, new trends in image processing are applied at first to general computer vision tasks, for example to 2D photos, before they are adapted and translated to tasks in the medical domain. This has several reasons. Firstly, 2D image processing is much less computationally intensive compared to processing large 3D image volumes from CTs or MRIs. Secondly, the algorithms are in general more "complex" and sophisticated (in terms of implementations) for 3D volumes than for 2D image processing, because they need to process one or more dimensions (if several scans at different time points have been acquired). Thirdly, often several image modalities and volumes, like combined positron emission tomography-computed tomography (PET-CT) scans, are available, and processing them jointly leads to information gain, but also increases the complexity. This is even more cumbersome if scans from different time points and/or different modalities, like CT and MRI, are not registered to each other. Finally, yet importantly, medical data is much harder to acquire and collect than for example natural images, especially in large quantities, not only because of the very time-consuming, often slice-by-slice manual ground truth generation and memory capacities, but also because of privacy concerns. Medical data is usually highly sensitive and personal, and therefore, using it for research purposes requires institutional review board (IRB) approvals and patient consent. Generally, data has to be pseudonymized / anonymized, by removing meta-information from the images and corresponding files, including name, sex and birth date. However, this is relatively easy compared to patient information that is encoded within the images themselves, like the patient's face in a head scan. Removing the eye area for patient de-identification within a 3D volume is possible, but laborious, because it must be done manually for every scan to make sure the volumes are properly de-identified. A fully automatic approach is conceivable, yet highly risky and potentially disastrous if it fails for even a single case. For head scans, de-identification by removing the eye area can be an option if the research is performed on a structure in another area of the head, like the lower jawbone [122,123], but on the downside, it can render the images unusable for applications requiring the entire volume, for example, facial-based medical augmented reality for the head and neck regions, for which all facial features are needed [124,125]. The IRB may allow the usage of the medical data for research purposes, but only within their own institution. This means that researchers from other institutions cannot re-use the existing data to validate published results or build upon existing methods to push the research boundaries. As a result, a considerable amount of effort has to be invested into obtaining IRB approval, acquiring data, and de-identification, which may delay new research by a few months, at best. Therefore, for large collections of rare pathological cases, it can easily take several years to establish a comprehensive database. Nonetheless, and against all odds, the massive amount of medical deep learning contributions is still increasing, and the proposed search strategy already reveals around 50 reviews or surveys for deep learning in PubMed by August 2020, which is more than all reviews from 2017 to 2019 together (Fig. 6). Equivalent to [58], we also looked at the locations of the first author's affiliations to get a sense of the geographical distribution of the medical deep learning reviews from this meta-review, and it reveals that the hotspots are the USA and China (Fig. 7). In case the first author provided several affiliations, we chose the very first one listed in the article.

The large amount of survey and review papers on medical deep learning published within the last three to four years is an indicator of the massive influence and importance that these algorithms already have in the medical community, and resulting clinical applications. This meta-review shows that, on average, a medical deep learning review has been published almost every month during the last years, with an approximately exponential increasing trend that seems to continue, if the distribution into the year 2020 is considered. Another indicator for the impact of deep learning in the medical field is the number of references (>5.000) and citations (>7.000) of the reviewed works. Besides the successes in outperforming state-of-the-art methods, there are several further reasons for and increase in research activities in (medical) deep learning:

- (1) The relatively easy application of deep learning algorithms to new data, enabled by comprehensive and user-friendly libraries and toolkits, like TensorFlow [126], PyTorch [127] or Caffe [128], just to name a few. These frameworks do not necessarily require an in-depth education in computer science. In contrast, in the era before deep learning, very good coding skills in programming languages like C or C++ were required to implement complex image processing algorithms. Factors like optimization for a reasonable runtime played a much larger role, as hardware was much weaker a few years ago.
- (2) Related to the first reason, most deep learning libraries and toolkits support Python bindings, which is a high-level, interpreted programming language, and therefore, easier to learn, apply and deploy compared to the aforementioned, compiled programming languages, like C or C++.
- (3) Relating to hardware, the broader availability of graphical processing units (GPUs) certainly contributed to the large distribution and application of medical deep learning and deep learning in general. Pretty much all deep learning libraries and toolkits natively support optimized training and execution of algorithms on a GPU, which speeds up the computation time many times over and makes many interesting big data applications possible. High-capacity GPUs decreased in price over the last few years, and GPU clusters, nowadays usually available at universities, research centers and companies, enable further parallelization and faster processing. Furthermore, GPU cloud servers and services (e.g. from Google Cloud or Amazon Web Services) can be accessed by everyone.
- (4) Another reason for the rapid spreading and adoption of deep learning (note that there is also already a review about artificial intelligence / deep learning techniques in imaging data acquisition, segmentation and diagnosis for covid-19 [129], and another one is on the horizon [130]), is that many researchers make their code publicly available to the research community, which is easily possible thanks to online repositories, like GitHub or GitLab. Because most implementations use common

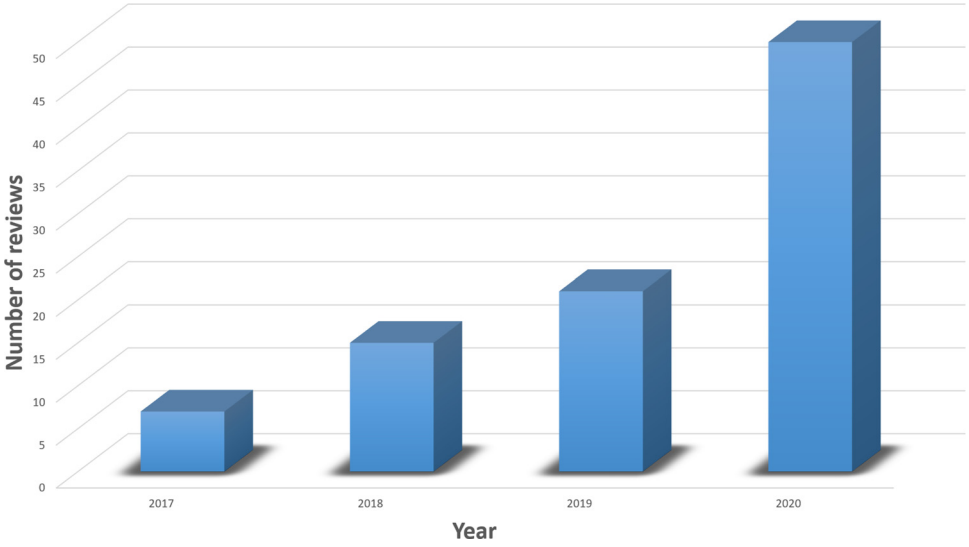

Fig. 6. Review and survey articles for medical deep learning in PubMed over the years (status as of August 2020).

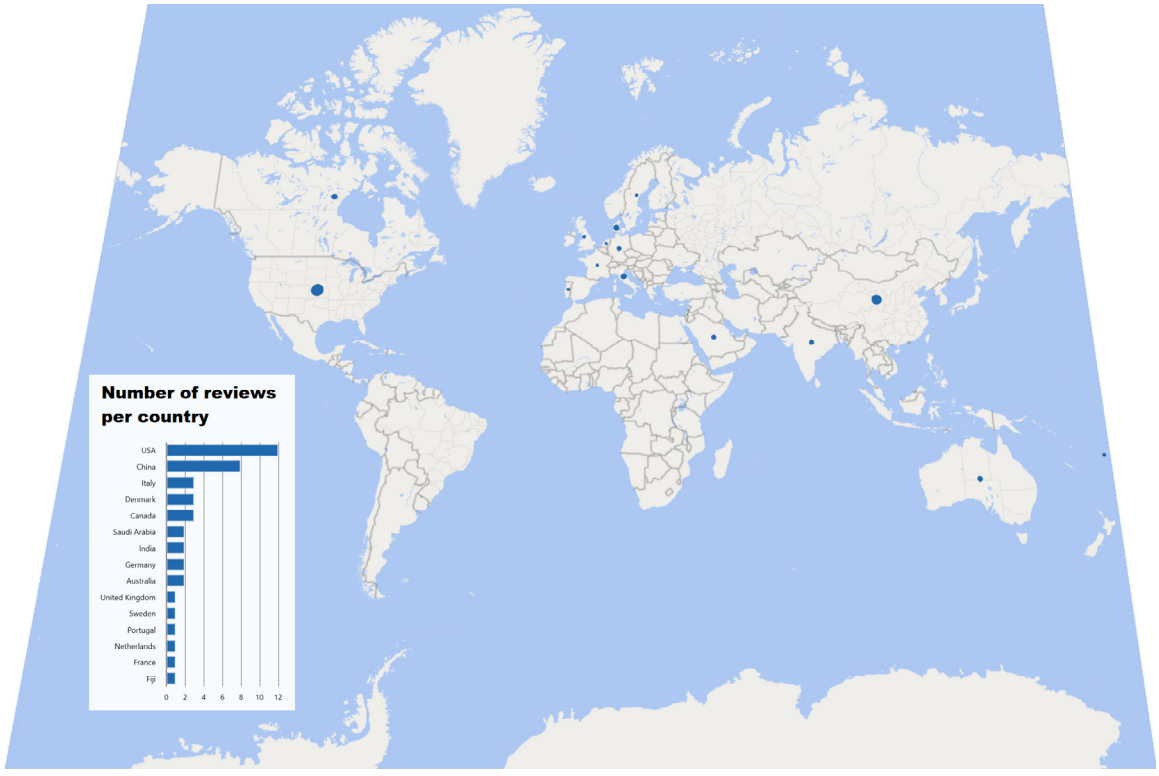

Fig. 7. World map showing the number of reviews per country according to the first author's affiliations.

deep learning toolkits, they can often be applied to new data without too much adaption.

- (5) The beforehand mentioned open access culture is promoted by publication venues, which require source code and data to be made openly available alongside the publication, like the Scientific Reports journal. This ensures reproducibility and verification by other researchers.
- (6) Furthermore, there are specific data journals, like Scientific Data or Data in Brief that provide venues to make medical datasets and data descriptors available to the research community. This makes it attractive to offer in-house datasets to the community (which is, first of all, a free service), because the data creators get an additional (citable) journal publication for their efforts.
- (7) Finally, deep learning is data-driven, which means it lives and dies by the amount of data it is fed, hence, the increasingly number of public medical databases, like the Cancer Imaging Archive or the Human Connectome Project, can be seen as very import driving forces behind the translation of deep learning into the medical domain.

It will be interesting to see what the future holds for us in the field of medical deep learning. Deep learning certainly already has an immense impact on the daily life of a large number of peo-

ple via the countless applications that are based on this technique, such as, virtual personal assistants like Amazon's Alexa, Apple's Siri or Google's Now. However, as several real-life examples recently demonstrated, deep learning algorithms are not inerrant, as evidenced by tragic car accidents with self-driving cars, racist missclassifications of images, or the machine learning bot Tay from Microsoft that became (some kind of) a (virtual) sexist neoNazi [131]. Another relevant example is Google Photos, which identified two black persons as gorillas back in 2015 [132]. Someone could argue that a human raised and educated in a sexist or racist environment might also develop a similar behavioral attitude: Whether the algorithm did this willingly is more of a philosophical discussion. Interestingly, Google "*fixed*" the problem by removing and blocking the image categories "*gorilla*", "*chimp*", "*chimpanzee*" and "*monkey*". So, in summary, even leading technology companies face difficulties when it comes to ensuring that the output of data-driven algorithms do not lead to prejudices, racism or stereotypes of any kind. If we translate this issue to the medical area, where complex 3D volumes are used for (life-critical) clinical support, this is very significant. It should also be mentioned that tasks where deep learning outperformed humans have often been performed under *laboratory conditions*, with a fixed set of samples, not including real-life tests, or further weaknesses [133], and recent publications show how deep neural networks can easily be fooled [134].

In summary, we identified the following primary research gaps while analyzing the reviewed works:

- Almost none of the reported deep learning algorithms were incorporated into clinical workflows, mostly due to ethics and trust concerns ("How can we trust the neural network not to be wrong/biased, when we don't understand why it answers the way it does?"), making the testing and integration into clinical practice a prominent research gap.
- Along the same vein, research into more interpretable "Explainable AI" constitutes a large research gap that is particularly relevant to understand the underlying methods. And even more relevant to healthcare is an evidence-based medicine where an efficacy must be demonstrated empirically [135].
- A lack of well-annotated, multi-institutional, public datasets (particularly for medical disciplines using data other than radiographic images) was reported by most review authors, who also suggested that many individual papers reported the potential for increased performance based on more data. This research gap still exists today (early 2022), with particular relevance in niche disciplines or concerning rare diseases, where the data volume is low to begin with, but decreases in significance over time, as more and more such datasets and other techniques, like Federated Learning [136], become available.
- There exists a distinct lack of reliable standardized key performance indicators for deep learning methods in the field of medical research. Therefore, standardization of data, data acquisition and performance reporting represents an important facet of deep learning (albeit less of a research gap and more of a trend in the field).
- The tuning of model architecture, data processing and augmentations, and training hyperparameter choice appears to have a significant effect on the eventual performance of the model. However, due to the "black box" nature of most deep learning models, optimal choices in this regard are often difficult to ascertain. Optimization of this trial-and-error process represents a significant research gap, which is already an intensively discussed topic in the wider deep learning community.
- Only a few works cover multimodal data and the majority of works focus on single-modality data. However, physicians consider a multitude of resources when treating patients, which computer-assisted methods should also do and there should be a stronger focus on methods that can simultaneously process multimodal data [137].

5. Author's perspective

From a high-level point of view, and to formulate it provocatively, some tasks like medical image segmentation have already been solved over thirty years ago, as can be seen by the claims within the countless publications released in the past years. In addition, the entire computer vision field seems to move from a general hot topic to another one over the years, like deformable models in the late '80 s [138], graph-based approaches in early '00 s [139], and, finally, deep neural networks after 2010 [140]. This is also reflected by the sharp drop or rise of citations for these publications, depending on the addressed methodology. A more realistic picture of the feasibilities of the proposed works during these times may be biomedical challenges, where authors are encouraged to develop algorithms for a specific task [141], for example the very influential brain tumor segmentation (BraTS) challenge, about the automatic segmentation of brain tumors or our new AutoImplant challenge from 2020 [142], about automatic cranial implant design. The quantitative and qualitative evaluation results are often presented afterwards in a compact summary publication [143]. This definitely enables a more objective view on what is currently possible with the state-of-the-art methods (in this regard, also note the new BIAS guidelines for transparent reporting of biomedical image analysis challenges [144]), even though such challenges usually cannot replace a real evaluation in a clinical setting.

Finally, it should be mentioned that most medical deep learning applications are still in an early phase of development and have not yet found their way into real clinical practice. This stands in strong contrast to non-learning approaches, like those used in medical navigation systems for neurosurgery [145,146]. However, most computer science venues for dissemination, especially flagship venues, explicitly prefer and demand new algorithms, while works that focus on the applicability of existing methods to real, variable, and noisy clinical scenarios are nipped in the bud with the argument that they lack technical novelty. At the same time, to foster their status in academia, researchers commonly need to fulfill the expectations of selected publication venues. In many situations, world-leading experts and members of the MICCAI community have been expressing concerns about the practical usability of the research output, too often limited to ideal scenarios. It is not uncommon to hear criticism about that fact that even high-impact conference proceedings usually contain a huge number of tools and algorithms that are designed for ideal or limited scenarios and may be therefore inapplicable or sometimes unneeded. MICCAI fellow D. Shen (author of the very first review article in the field of medical deep learning according to our search strategy, see epub date in Table 2) summed up this issue in a recent public statement on LinkedIn [147]: "*In MICCAI field, people are studying same problems (sometimes ideal problems) with very similar methods for many years. Everyone claims their method is new (although mostly just simply borrowing from others). This is very serious issue, since people in this small academic field judge contributions of their works by themselves. If MICCAI people can just move a little bit out of their academic field, i.e., thinking more on real applications in clinical workflow, this issue can be largely avoided. We, as faculty, have more responsibility for changing this situation*". A step towards this direction could be that interdisciplinary and application-oriented venues encourage the involvement of a medical partner, including a statement of feasibility in the clinical practice. Furthermore, several interdisciplinary venues do not explicitly require any IRB approval statement, even if the manuscripts deal with clinical patient data (an exception here are publicly available datasets, but

many works are still evaluated on private datasets provided by a medical partner, which also hinders an objective analyzing and reporting in reviews). In most medical venues, this is a standard requirement and submissions are rejected if the manuscript does not contain an official IRB or patient consent statement. Note that an approval from an ethics commission is also a pre-check for reasonability and should stop research endeavors that would harm the patient, for example by additional radiation exposure, or do not adhere to clinical workflows.

Nevertheless, and to pick up the "*not yet found their way into the real clinical practice*" and "*limited to ideal scenarios*" thoughts from before, some deep learning experts claim that just adding enough (training) data will automatically lead to perfect results. Contradictory to this opinion, cars have driven already millions of kilometers to acquire training data, but fully self-driving capabilities are still far away from being reliable, especially under different weather and light conditions. In this light, Tesla recently removed the "*full self-driving*" option from its car store on its website and Uber completely abandoned the development of self-driving cars. It should also be noted that in the medical field, such a massive amount of data will, in many cases, never be available. Certain pathologies are simply (and thankfully) not frequent enough, so even by collecting all the patient data for this pathology from the hospitals around the world and applying additional data augmentation methods [148], there still might not be enough for training powerful algorithms.

Nonetheless, there have been certain tasks where machine learning has undoubtedly outperformed humans already. Examples are Deep Blue [149] and AlphaGo [150] in games, where machine learning algorithms could even beat the best (known) human players around the world. However, these tasks have strong constraints by fixes rules on which algorithms can rely on. In contrast, medical tasks usually do not follow such rules and theoretically, unlimited possibilities exist. For example, a brain tumor [151] looks different for every patient in terms of shape, size, texture, etc. Another example is the human voice, with individual pitches and pronunciations, and further the inter-human variations when expressing different emotions [152]. In addition, algorithms can fall back on a massive database of pre-trained games and game moves, without any further uncertainty. Another example where deep learning works very well in practice is the automatic detection and analysis of car licenses. Despite several challenges and uncertainties, like different fonts, colors, languages, deformities, complex backgrounds, hazardous situations, speeding vehicles, occlusion, horizontal or vertical skew, blurriness, and illumination diversions [153], the recognition task still stays within a restricted rule set. Therefore, learning algorithms can be pre-trained, for example, by just going over the alphabet with variations, like changing the font, colors, adding some occlusion, etc. It should be kept in mind that still, vehicular license plate recognition is far from perfect.

In principle, deep learning is trying to mimic the human brain, especially the learning process of a human brain [154]. Equivalent to the fact that we cannot look into someone's brain with its thoughts or mindset, it is also not yet fully understood what is going on inside a deep neural network (even though we have access to all neurons and its connections, in contrast to a human's brain) [155]. Hence, it is as hard to predict exceptions and failures as seen in recent events, like car accidents, as it is to foresee human behavior and mistakes (even if there are, ironically, deep learning works that try to predict human behavior [156]). Trained neural networks with several layers and with a few hundred or a few thousand neurons are not understandable anymore in all detail [155]. This stands in strong contrast to pure engineering approaches, which can be understood in every detail. That makes the acceptance of such black box (some even call it Voodoo [157]) approaches, like deep learning, by the general population much harder. At this point, we want to refer the interested reader to the concept of disentanglement, which tries to make latent representations interpretable [158].

To conclude, deep learning is an exciting new field with a lot of potential, but not free of controversies. We believe that this first meta-review of medical deep learning reviews and surveys can provide a quick and comprehensive reference for scientists (or just interested readers) who want to get a high-level overview of this field, and maybe want to contribute and thus, accelerate the development in medical deep learning. Hence, the contribution of this systematic meta-review is sixfold:

- providing an overview of current deep learning reviews where a medical application plays the key role,
- arranging the researched works chronologically for a historical "*roten Faden*" (*red/common thread*) and picture over the years,
- extracting the overall number of referenced works and citations to give an impression of the research influence and footprints of the respective field,
- analyzing, exploring and highlighting the main reasons for the massive research efforts on this topic,
- conducting a comprehensive discussion of the current state-of-the-art methods in the deep learning area with achievements but also failures from other domains that should be avoided in the medical area,
- and providing a critical expert opinion and pointing out further controversies.

Declaration of Competing Interest

The authors declare no competing financial interests.

Acknowledgements

This work received funding from the Austrian Science Fund (FWF) KLI 678-B31: "*enFaced: Virtual and Augmented Reality Training and Navigation Module for 3D-Printed Facial Defect Reconstructions*", FWF KLI 1044: "*Instant AR Tool for Maxillofacial Surgery*" and the TU Graz Lead Project (*Mechanics, Modeling and Simulation of Aortic Dissection*). Moreover, this work was supported by CAMed (COMET K-Project 871132), which is funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT), and the Austrian Federal Ministry for Digital and Economic Affairs (BMDW), and the Styrian Business Promotion Agency (SFG). Furthermore, we acknowledge the REACT-EU project KITE (Plattform für KI-Translation Essen). Finally, we want to make the interested reader aware of our medical image processing framework *StudierFenster* (www.studierfenster.at) [159], where medical deep learning approaches can be tried out in a standard web browser.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2022.106874.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015 May) 436–444.
[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5505–5514.
[3] B. Liu, J. Liu, Overview of Image Denoising Based on Deep Learning, Journal of Physics: Conference Series, 1176, IOP Publishing, 2019 Mar.
[4] A. Loquercio, M. Segu, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, IEEE Robot. Automat. Lett. 5 (2) (2020 Feb 18) 3153–3160.
[5] H. Fujiyoshi, T. Hirakawa, T. Yamashita, Deep learning-based image recognition for autonomous driving, IATSS Res. 43 (4) (2019 Dec 1) 244–252.

[6] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[7] F. Fuchs, Y. Song, E. Kaufmann, D. Scaramuzza, P. Duerr, Super-Human Performance in Gran Turismo Sport Using Deep Reinforcement Learning. arXiv preprint arXiv:2008.07971. 2020 Aug 18.

[8] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016 Jan) 484–489.

[9] S. Dash, S.K. Shakyawar, M. Sharma, S. Kaushik, Big data in healthcare: management, analysis and future prospects, J. Big Data 6 (1) (2019 Dec 1) 54.

[10] M. Franceschet, The role of conference publications in CS, Commun. ACM 53 (12) (2010 Dec 1) 129–132.

[11] M. Eckmann, A. Rocha, J. Wainer, Relationship between high-quality journals and conferences in computer vision, Scientometrics 90 (2) (2012 Feb 1) 617–630.

[12] A.V. Dalca, J. Guttag, M.R. Sabuncu, Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9290–9299.

[13] A. Pepe, J. Li, M. Rolf-Pissarczyk, C. Gsaxner, X. Chen, G.A. Holzapfel, J. Egger, Detection, segmentation, simulation and visualization of aortic dissections: a review, Med. Image Anal. (2020 Jul 7) 101773.

[14] E. Ernst, A systematic review of systematic reviews of homeopathy, Br. J. Clin. Pharmacol. 54 (6) (2002 Dec) 577–582.

[15] A. Chatzimparmpas, R.M. Martins, I. Jusufi, A. Kerren, A survey of surveys on the use of visualization for interpreting machine learning models, Inf. Vis. (2020 Mar 19) 1473871620904671.

[16] H. Liang, X. Sun, Y. Sun, Y. Gao, Text feature extraction based on deep learning: a review, EURASIP J. Wirel. Commun. Netw. 2017 (1) (2017 Dec) 1–2.

[17] F. Hohman, M. Kahng, R. Pienta, D.H. Chau, Visual analytics in deep learning: an interrogative survey for the next frontiers, IEEE Trans. Vis. Comput. Graph. 25 (8) (2018 Jun 4) 2674–2693.

[18] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018 Feb 1).

[19] Q. Zhao, P. Kong, J. Min, Y. Zhou, Z. Liang, S. Chen, M. Li, A review of deep learning methods for the detection and classification of pulmonary nodules, J. Biomed. Eng. 36 (6) (2019 Dec 1) 1060–1068.

[20] Y. Liu, Z. Zhao, Review of research on detection and tracking of minimally invasive surgical tools based on deep learning, J. Biomed. Eng. 36 (5) (2019 Oct) 870–878.

[21] K.A. Weigel, P.M. VanRaden, H.D. Norman, H. Grosu, A 100-Year Review: methods and impact of genetic selection in dairy cattle—From daughter–dam comparisons to deep learning algorithms, J. Dairy Sci. 100 (12) (2017 Dec 1) 10234–10250.

[22] L. Cadorin, A. Bagnasco, A. Tolotti, N. Pagnucci, L. Sasso, Instruments for measuring meaningful learning in healthcare students: a systematic psychometric review, J. Adv. Nurs. 72 (9) (2016 Sep) 1972–1990.

[23] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006 Jul) 1527–1554.

[24] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006 Jul 28) 504–507.

[25] J. Biggs, D. Kember, D.Y. Leung, The revised two-factor study process questionnaire: R-SPQ-2F, British J. Edu. Psychol. 71 (1) (2001 Mar) 133–149.

[26] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015 Jan 1) 85–117.

[27] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarenkov, A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion (2021 May 23).

[28] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017 Jun 21) 221–248.

[29] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2018 Nov) 1236–1246.

[30] G. Litjens, T. Kooi, B.E. Bejnordi, A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017 Dec 1) 60–88.

[31] R. Feng, M. Badgeley, J. Mocco, E.K. Oermann, Deep learning guided stroke management: a review of clinical applications, J. Neurointerv. Surg. 10 (4) (2018 Apr 1) 358–362.

[32] Y. Xue, S. Chen, J. Qin, Y. Liu, B. Huang, H. Chen, Application of deep learning in automated analysis of molecular images in cancer: a survey, Contrast. Media Mol. Imaging 2017 (2017 Oct 15).

[33] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, IEEE J. Biomed. Health Inform. 22 (5) (2017 Oct 27) 1589–1604.

[34] F. Xing, Y. Xie, H. Su, F. Liu, L. Yang, Deep learning in microscopy image analysis: a survey, IEEE Trans. Neural Netw. Learn. Syst. 29 (10) (2017 Nov 22) 4550–4568.

[35] W. Tang, J. Chen, Z. Wang, H. Xie, H. Hong, Deep learning for predicting toxicity of chemicals: a mini review, J. Environ. Sci. Health, Part C 36 (4) (2018 Oct 2) 252–271.

[36] Y. Yang, X. Feng, W. Chi, Z. Li, W. Duan, H. Liu, W. Liang, W. Wang, P. Chen, J. He, B Liu, Deep learning aided decision support for pulmonary nodules diagnosing: a review, J. Thorac. Dis. 10 (Suppl 7) (2018 Apr) S867.

[37] O. Faust, Y. Hagiwara, T.J. Hong, O.S. Lih, U.R. Acharya, Deep learning for healthcare applications based on physiological signals: a review, Comput. Methods Programs Biomed. 161 (2018 Jul 1) 1–3.

[38] F. Celesti, A. Celesti, J. Wan, M. Villari, Why deep learning is changing the way to approach NGS data processing: a review, IEEE Rev. Biomed. Eng. 11 (2018 Apr 12) 68–76.

[39] P. Meyer, V. Noblet, C. Mazzara, A. Lallement, Survey on deep learning for radiotherapy, Comput. Biol. Med. 98 (2018 Jul 1) 126–146.

[40] P.S. Grewal, F. Oloumi, U. Rubin, M.T. Tennant, Deep learning in ophthalmology: a review, Can. J. Ophthalmol. 53 (4) (2018 Aug 1) 309–313.

[41] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018 Oct) 1419–1428.

[42] K. Lan, D.T. Wang, S. Fong, L.S. Liu, K.K. Wong, N. Dey, A survey of data mining and deep learning in bioinformatics, J. Med. Syst. 42 (8) (2018 Aug 1) 139.

[43] S. Zhang, S.M. Bamakan, Q. Qu, S. Li, Learning for personalized medicine: a comprehensive review from a deep learning perspective, IEEE Rev. Biomed. Eng. 12 (2018 Aug 7) 194–208.

[44] N. Ganapathy, R. Swaminathan, T.M. Deserno, Deep learning on 1-D biosignals: a taxonomy-based survey, Yearb. Med. Inform. 27 (1) (2018 Aug) 98.

[45] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, S. Peng, Deep learning in omics: a survey and guideline, Brief Funct. Genomics 18 (1) (2019 Jan) 41–57.

[46] E.E. Cust, A.J. Sweeting, K. Ball, S. Robertson, Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance, J. Sports Sci. 37 (5) (2019 Mar 4) 568–600.

[47] K.B. Nielsen, M.L. Lautrup, J.K. Andersen, T.R. Savarimuthu, J. Grauslund, Deep learning–based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance, Ophthalmol. Retina 3 (4) (2019 Apr 1) 294–304.

[48] A. Gupta, P.J. Harrison, H. Wieslander, N. Pielawski, K. Kartasalo, G. Partel, L. Solorzano, A. Suveer, A.H. Klemm, O. Spjuth, I.M. Sintorn, Deep learning in image cytometry: a review, Cytometry Part A 95 (4) (2019 Apr) 366–380.

[49] M.A. Mazurowski, M. Buda, A. Saha, M.R. Bashir, Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI, J. Magnet. Reson. Imag. 49 (4) (2019 Apr) 939–954.

[50] M. Biswas, V. Kuppili, L. Saba, D.R. Edla, H.S. Suri, E. Cuadrado-Godia, J.R. Laird, R.T. Marinhoe, J.M. Sanches, A. Nicolaides, J.S. Suri, State-of-the-art review on deep learning in medical imaging, Front. Biosci. (Landmark Ed) 24 (2019 Jan) 392–426.

[51] G.S. Tandel, M. Biswas, O.G. Kakde, A. Tiwari, H.S. Suri, M. Turk, J.R. Laird, C.K. Asare, A.A. Ankrah, N.N. Khanna, B.K. Madhusudhan, A review on a deep learning perspective in brain cancer classification, Cancers (Basel) 11 (1) (2019 Jan) 111.

[52] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review, J. Neural Eng. 16 (3) (2019 Apr 9) 031001.

[53] L.M. Pehrson, M.B. Nielsen, Ammitzbøl Lauridsen C. Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: a systematic review, Diagnostics 9 (1) (2019 Mar) 29.

[54] M.M. Shaver, P.A. Kohanteb, C. Chiou, M.D. Bardis, C. Chantaduly, D. Bota, C.G. Filippi, B. Weinberg, J. Grinband, D.S. Chow, P.D. Chang, Optimizing neuro-oncology imaging: a review of deep learning approaches for glioma imaging, Cancers (Basel) 11 (6) (2019 Jun) 829.

[55] N. Asiri, M. Hussain, F. Al Adel, N Alzaidi, Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey, Artif. Intell. Med. 99 (2019 Aug 1) 101701.

[56] S. Parvaneh, J. Rubin, S. Babaeizadeh, M. Xu-Wilson, Cardiac arrhythmia detection using deep learning: a review, J. Electrocardiol. 57 (2019 Nov 1) S70–S74.

[57] W. Wardah, M.G. Khan, A. Sharma, M.A. Rashid, Protein secondary structure prediction using neural networks and deep learning: a review, Comput. Biol. Chem. 81 (2019 Aug 1) 1–8.

[58] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019 Aug 14) 051001.

[59] A.A. Valliani, D. Ranti, E.K. Oermann, Deep learning and neurology: a systematic review, Neurol. Ther. (2019 Aug 21) 1–5.

[60] K. Munir, H. Elahi, A. Ayub, F. Frezza, A. Rizzi, Cancer diagnosis using deep learning: a bibliographic review, Cancers (Basel) 11 (9) (2019 Sep) 1235.

[61] Z. Akkus, J. Cai, A. Boonrod, A. Zeinoddini, A.D. Weston, K.A. Philbrick, B.J. Erickson, A survey of deep-learning applications in ultrasound: artificial intelligence–powered ultrasound for improving clinical workflow, J. Am. College Radiol. 16 (9) (2019 Sep 1) 1318–1328.

[62] L. Boldrini, J.E. Bibault, C. Masciocchi, Y. Shen, M.I. Bittner, Deep learning: a review for the radiation oncologist, Front. Oncol. 9 (2019) 977.

[63] T. Zhang, J. Leng, Y. Liu, Deep learning for drug–drug interaction extraction from the literature: a review, Brief. Bioinform. (2019 Nov 4).

[64] R. Suarez-Ibarrola, S. Hein, G. Reis, C. Gratzke, A. Miernik, Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer, World J. Urol. (2019 Nov 5) 1–9.

[65] S.S. Mostafa, F. Mendonça, A. G Ravelo-García, F Morgado-Dias, A systematic review of detecting sleep apnea using deep learning, Sensors 19 (22) (2019 Jan) 4934.

[66] S. Sengupta, A. Singh, H.A. Leopold, T. Gulati, V. Lakshminarayanan, Ophthalmic diagnosis using deep learning with fundus images–A critical review, Artif. Intell. Med. 102 (2020 Jan 1) 101758.

[67] M.A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review, Comput. Methods Programs Biomed. 187 (2020 Apr 1) 105242.

[68] D. Li, B. Mikela Vilmun, J. Frederik Carlsen, E. Albrecht-Beste, C. Ammitzbøl Lauridsen, M. Bachmann Nielsen, K. Lindskov Hansen, The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from lidc-idri: a systematic review, Diagnostics 9 (4) (2019 Dec) 207.

[69] S.A. Azer, Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: a systematic review, World. J. Gastrointest. Oncol. 11 (12) (2019 Dec 15) 1218.

[70] J. Ma, Y. Song, X. Tian, Y. Hua, R. Zhang, J. Wu, Survey on deep learning for pulmonary medical imaging, Front. Med. (2019 Dec 16) 1–20.

[71] K. Fukushima, S. Miyake, Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and Cooperation in Neural Nets 1982 (pp. 267–285). Springer, Berlin, Heidelberg.

[72] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998 Nov) 2278–2324.

[73] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems 2012 (pp. 1097–1105).

[74] J. Egger, A. Pepe, C. Gsaxner, Y. Jin, J. Li, R Kern, Deep learning—A first meta–survey of selected reviews across scientific disciplines, their commonalities, challenges and research impact, Peer J. Comput. Sci. 7 (2021 Nov 17) e773.

[75] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2) (2020 Feb) 261–318.

[76] Z.Q. Zhao, P. Zheng, S.T. Xu, X. Wu, Object detection with deep learning: a review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019 Jan 28) 3212–3232.

[77] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, R. Qu, A survey of deep learning-based object detection, IEEE Access 7 (2019 Sep 5) 128837–128868.

[78] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez–Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. 70 (2018 Sep 1) 41–65.

[79] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey. arXiv preprint arXiv:2001.05566. 2020 Jan 15.

[80] I. Masi, Y. Wu, T. Hassner, P. Natarajan, Deep face recognition: a survey, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2018 Oct 29, pp. 471–478.

[81] S. Li, W. Deng, Deep facial expression recognition: a survey, IEEE Trans. Affect. Comput. (2020 Mar 17).

[82] W. Mei, W. Deng, Deep face recognition: a survey. arXiv preprint arXiv:1804.06655. 2018;1.

[83] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vis. Comput. 60 (2017 Apr 1) 4–21.

[84] P. Wang, W. Li, P. Ogunbona, J. Wan, S. Escalera, RGB-D-based human motion recognition with deep learning: a survey, Comput. Vis. Image Understand 171 (2018 Jun 1) 118–139.

[85] K. Sundararajan, D.L. Woodard, Deep learning for biometrics: a survey, ACM Comput. Surv. (CSUR) 51 (3) (2018 May 23) 1–34.

[86] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, Biometric recognition using deep learning: a survey. arXiv preprint arXiv:1912.00271. 2019 Nov 30.

[87] Z. Wang, J. Chen, S.C. Hoi, Deep learning for image super-resolution: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2020 Mar 23).

[88] M.Z. Hossain, F. Sohel, M.F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Comput. Surv. (CSUR) 51 (6) (2019 Feb 4) 1–36.

[89] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big. Data 6 (1) (2019 Dec 1) 60.

[90] Z. Wang, Q. She, T.E. Ward, Generative adversarial networks in computer vision: a survey and taxonomy. arXiv preprint arXiv:1906.01529. 2019 Jun 4.

[91] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Mag. 13 (3) (2018 Jul 20) 55–75.

[92] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: core tasks, applications and evaluation, J. Artif. Intell. Res. 61 (2018 Jan 27) 65–170.

[93] S. Santhanam, S. Shaikh, A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. arXiv preprint arXiv:1906.00500. 2019 Jun 2.

[94] J. Gao, M. Galley, L. Li, Neural approaches to conversational AI, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018 Jun 27, pp. 1371–1374.

[95] H. Chen, X. Liu, D. Yin, J. Tang, A survey on dialogue systems: recent advances and new frontiers, ACM Sigkdd Expl. Newslett. 19 (2) (2017 Nov 21) 25–35.

[96] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. (2020 Mar 17).

[97] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470. 2019 Oct 25.

[98] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey, Wiley Interdis. Rev. 8 (4) (2018 Jul) e1253.

[99] H.H. Do, P.W. Prasad, A. Maag, A. Alsadoon, Deep learning for aspect-based sentiment analysis: a comparative review, Expert Syst. Appl. 118 (2019 Mar 15) 272–299.

[100] T. Shi, Y. Keneshloo, N. Ramakrishnan, C.K. Reddy, Neural abstractive text summarization with sequence-to-sequence models. arXiv preprint arXiv:1812.02303. 2018 Dec 5.

[101] T. Lai, T. Bui, S. Li, A review on deep learning techniques applied to answer selection, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018 Aug, pp. 2132–2144.

[102] Y. Zhang, M.M. Rahman, A. Braylan, B. Dang, H.L. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, Neural information retrieval: a literature review. arXiv preprint arXiv:1611.06792. 2016 Nov 18.

[103] F. Almeida, G. Xexéo, Word embeddings: a survey. arXiv preprint arXiv:1901.09069. 2019 Jan 25.

[104] F.Z. Xing, E. Cambria, R.E. Welsch, Natural language based financial forecasting: a survey, Artif. Intell. Rev. 50 (1) (2018 Jun 1) 49–73.

[105] Q. Zhang, L.T. Yang, Z. Chen, P. Li, A survey on deep learning for big data, Inf. Fus. 42 (2018 Jul 1) 146–157.

[106] M. Mohammadi, A. Al-Fuqaha, S. Sorour, M. Guizani, Deep learning for IoT big data and streaming analytics: a survey, IEEE Commun. Surv. Tutor. 20 (4) (2018 Jun 6) 2923–2960.

[107] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M Dehmer, An introductory review of deep learning for prediction models with big data, Front. Artif. Intell. 3 (2020) 4.

[108] S.S. Mousavi, M. Schukat, E. Howley, Deep reinforcement learning: an overview. In Proceedings of SAI Intelligent Systems Conference 2016 Sep 21 (pp. 426–440). Springer, Cham.

[109] Y. Li, Deep reinforcement learning: an overview. arXiv preprint arXiv:1701.07274. 2017 Jan 25.

[110] K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath, Deep reinforcement learning: a brief survey, IEEE Signal Process. Mag. 34 (6) (2017 Nov 9) 26–38.

[111] C. Zhang, P. Patras, H. Haddadi, Deep learning in mobile and wireless networking: a survey, IEEE Commun. Surv. Tutor. 21 (3) (2019 Mar 13) 2224–2287.

[112] K. Ota, M.S. Dao, V. Mezaris, F.G. Natale, Deep learning for mobile multimedia: a survey, ACM Trans. Multim. Comput. Commun. Appl. (TOMM 13 (3s) (2017 Jun 28) 1–22.

[113] D. Ramachandram, G.W. Taylor, Deep multimodal learning: a survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017 Nov 9) 96–108.

[114] J.E. Ball, D.T. Anderson, C.S. Chan, Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community, J. Appl. Remote Sens. 11 (4) (2017 Sep) 042609.

[115] Z. Zhang, P. Cui, W. Zhu, Deep learning on graphs: a survey, IEEE Trans. Knowl. Data Eng. (2020 Mar 17).

[116] D. Kwon, H. Kim, J. Kim, S.C. Suh, I. Kim, K.J. Kim, A survey of deep learning-based network anomaly detection, Cluster Comput. (2019 Jan) 1–3.

[117] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: a survey and new perspectives, ACM Comput. Surv. (CSUR) 52 (1) (2019 Feb 25) 1–38.

[118] A. Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agricul. 147 (2018 Apr 1) 70–90.

[119] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M.P. Reyes, M.L. Shyu, S.C. Chen, S.S. Iyengar, A survey on deep learning: algorithms, techniques, and applications, ACM Comput. Surv. (CSUR) 51 (5) (2018 Sep 18) 1–36.

[120] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning, Arch. Comput. Methods Eng. (2019 Jun 1) 1–22.

[121] M. Raghu, E. Schmidt, A survey of deep learning for scientific discovery. arXiv preprint arXiv:2003.11755. 2020 Mar 26.

[122] J. Wallner, M. Schwaiger, K. Hochegger, C. Gsaxner, W. Zemann, J. Egger, A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery, Comput. Methods Programs Biomed. 182 (2019 Dec 1) 105102.

[123] J. Wallner, I. Mischak, J. Egger, Computed tomography data collection of the complete human mandible and valid clinical ground truth models, Sci. Data 6 (2019 Jan 29) 190003.

[124] C. Gsaxner, A. Pepe, J. Wallner, D. Schmalstieg, J. Egger, Markerless image-to-face registration for untethered augmented reality in head and neck surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, Springer, 2019 Oct 13, pp. 236–244.

[125] C. Gsaxner, J. Wallner, X. Chen, W. Zemann, J. Egger, Facial model collection for medical augmented reality in oncologic cranio-maxillofacial surgery, Sci. Data 6 (1) (2019 Dec 9) 1–7.

[126] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, Tensorflow: a system for large-scale machine learning, in: 12th {USENIX} Symposium On Operating Systems Design And Implementation ({OSDI} 16), 2016, pp. 265–283.

[127] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch.

[128] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014 Nov 3, pp. 675–678.

[129] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19, IEEE Rev. Biomed. Eng. (2020 Apr 16).
[130] M. Islam, F. Karray, R. Alhajj, J. Zeng, A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). arXiv preprint arXiv:2008.04815. 2020 Aug 9.
[131] A. Shuldiner, in: Raising Them Right: AI and the Internet of Big Things. In Artificial Intelligence For the Internet of Everything, Academic Press, 2019 Jan 1, pp. 139–143.
[132] R. Yu, G.S. Alì, What's inside the Black Box? AI Challenges for Lawyers and Researchers, Legal Inf. Manage. 19 (1) (2019 Mar) 2–13.
[133] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Health 1 (6) (2019 Oct 1) e271–e297.
[134] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: high confidence predictions for unrecognizable images, in: InProceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 427–436.
[135] L.G. McCoy, C.T. Brenna, S.S. Chen, K. Vold, S. Das, Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based, J. Clin. Epidemiol. 142 (2022 Feb 1) 252–257.
[136] N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, S. Ourselin, The future of digital health with federated learning, NPJ Digit. Med. 3 (1) (2020 Sep 14) 1–7.
[137] L. Heiliger, A. Sekuboyina, B. Menze, J. Egger, J. Kleesiek, Beyond medical imaging-a review of multimodal deep learning in radiology.
[138] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, Int. J. Comput. Vis. 1 (4) (1988 Jan 1) 321–331.
[139] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001 Nov) 1222–1239.
[140] Q.V. Le, Building high-level features using large scale unsupervised learning, in: 2013 IEEE International Conference On Acoustics, Speech and Signal Processing, IEEE, 2013 May 26, pp. 8595–8598.
[141] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A.P. Bradley, A. Carass, C. Feldmann, Why rankings of biomedical image analysis competitions should be interpreted with care, Nat. Commun. 9 (1) (2018 Dec 6) 1–3.
[142] J. Li, J. Egger, in: Towards the Automatization of Cranial Implant Design in Cranioplasty. First Challenge, AutoImplant 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings, Springer Nature, 2020.
[143] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014 Dec 4) 1993–2024.
[144] L. Maier-Hein, A. Reinke, M. Kozubek, A.L. Martel, T. Arbel, M. Eisenmann, A. Hanbury, P. Jannin, H. Müller, S. Onogur, J. Saez-Rodriguez, BIAS: transparent reporting of biomedical image analysis challenges, Med. Image Anal. 66 (2020 Dec 1) 101796.
[145] C. Nimsky, O. Ganslandt, B. von Keller, J. Romstöck, R. Fahlbusch, Intraoperative high-field-strength MR imaging: implementation and experience in 200 patients, Radiology 233 (1) (2004 Oct) 67–78.
[146] J.S. Perlmutter, J.W. Mink, Deep brain stimulation, Annu. Rev. Neurosci. 29 (2006 Jul 21) 229–257.
[147] D. Shen, Public Statement. LinkedIn. 2020 Oct (accessed on 11/24/2020). https://www.linkedin.com/feed/update/urn:li:activity:6719177936513089536/
[148] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification, in: 2018 IEEE 15th International Symposium On Biomedical Imaging (ISBI 2018), IEEE, 2018 Apr 4, pp. 289–293.
[149] M. Campbell, A.J. Hoane Jr, F.H Hsu, Deep blue, Artif. Intell. 134 (1–2) (2002 Jan 1) 57–83.
[150] J.X. Chen, The evolution of computing: AlphaGo, Comput. Sci. Eng. 18 (4) (2016 Jul) 4–7.
[151] J. Egger, T. Kapur, A. Fedorov, S. Pieper, J.V. Miller, H. Veeraraghavan, B. Freisleben, A.J. Golby, C. Nimsky, R. Kikinis, GBM volumetry using the 3D Slicer medical image computing platform, Sci. Rep. 3 (1) (2013 Mar 4) 1–7.
[152] S. Johar, Emotion, Affect and Personality in Speech: The Bias of Language and Paralanguage, Springer, 2015 Dec 22.
[153] M.Y. Arafat, A.S. Khairuddin, U. Khairuddin, R. Paramesran, Systematic review on vehicular licence plate recognition framework in intelligent transport systems, IET Intell. Transp. Syst. 13 (5) (2019 Jan 2) 745–755.
[154] V. C̄inivāca Cakkaravartti, Demystifying the Brain: A Computational Approach, Springer, 2019.
[155] V. Buhrmester, D. Münch, M. Arens, Analysis of explainers of black box deep neural networks for computer vision: a survey. arXiv preprint arXiv:1911.12116. 2019 Nov 27.
[156] J.S. Hartford, J.R. Wright, K. Leyton-Brown, Deep learning for predicting human strategic behavior. In Advances in Neural Information Processing Systems 2016 (pp. 2424–2432).
[157] S. Saeb, L. Lonini, A. Jayaraman, D.C. Mohr, K.P. Kording, Voodoo machine learning for clinical predictions, bioRxiv (2016 Jan 1) 059774.
[158] J. Fragemann, L. Ardizzone, J. Egger, J. Kleesiek, Review of Disentanglement Approaches for Medical Applications–Towards Solving the Gordian Knot of Generative Models in Healthcare. arXiv preprint arXiv:2203.11132. 2022 Mar 21.
[159] J. Egger, D. Wild, M. Weber, C.A. Bedoya, F. Karner, A. Prutsch, M. Schmied, C. Dionysio, D. Krobath, Y. Jin, C. Gsaxner, Studierfenster: an open science cloud-based medical imaging analysis platform, J. Digit. Imaging (2022 Jan 21) 1–6.

1 Supplemental Material

The purpose of this section is to give the interested reader a 'lightweight' introduction about the technical background of deep learning, for a better understanding of our meta-review contribution [1] and to make the manuscript self-contained. Thus, we avoid that the reader of our manuscript needs to consult other, additional sources for an overall understanding of our contribution. In doing so, we stay mostly within the 'general' deep learning domain, not explicitly focusing only on medical-specific concepts, because most deep learning-based concepts studied in the medical domain originate from classic non-medical domains, like computer vision, and can be used and applied to several domains. Only at the end of this supplemental material, we discuss some segmentation characteristics for medical images. Readers who are already familiar with the basic concepts of deep learning, can skip this section and dive directly into the main body of our manuscript.

The first section, 1.1, of this supplemental material, explains the basic principles and structures of artificial neural networks (ANNs) and how they work, beginning from a single perceptron to multilayer perceptrons and common activation functions. Then, in section 1.2 more detailed information about deep neural networks is given, in which the process of training is explained, including loss functions as well as different parameter optimizers, and common issues of training ANNs. Next, insights into the theory of Convolutional neural networks (CNNs), autoencoders (AE) and variational autoencoders (VAE) are stated. Then, an overview of different methods of image segmentation is listed in section 1.3. Furthermore, an overview of PyTorch, a common deep learning library used by the research community, is provided in section 1.4. This section closes with basic information about CUDA in section 1.5, the platform developed by Nvidia, which is commonly used to accelerate the training process of ANNs.

1.1 Artificial Neural Networks

The basic architecture of artificial neural networks (ANNs) is inspired in a simplified way by the processes of the human brain. A description of the neural activity is provided by McCulloch and Pitts [2]. ANNs are artificial systems, which work in an adaptive manner and can modify their internal relations and structures. Neural networks are particularly suitable for tasks in the nonlinear domain. In particular, during the process of training ANNs, these try to understand the problems and rules of the task [3].

1.1.1 Perceptron

The simplest type of ANN is called perceptron (Figure 1.1) and was designed by Rosenblatt [4]. The perceptron is a single computational layered network and consists of an input and output layer. The output y, as stated in equation 1.1, is calculated by summing up the n input features of the input vector $\boldsymbol{X} = [x_1, x_2, ..., x_n]$ times the corresponding weights $\boldsymbol{W} = [w_1, w_2, ..., w_n]$ and passing the result into an activation function *f(·)*. In some cases, a bias b is added to the sum of weighted feature values, which represents the invariant part of the prediction and is stated in equation 1.2 [5].

$$y = f(\boldsymbol{X} \cdot \boldsymbol{W}) = f(\sum_{i=1}^{n} x_i w_i) \tag{1.1}$$

$$y = f(\boldsymbol{X} \cdot \boldsymbol{W} + b) = f(\sum_{i=1}^{n} x_i w_i + b) \tag{1.2}$$

The perceptron algorithm uses the *sign* function (1.3) as an activation function to perform binary linear classification on the input data (1.4). The sum of weighted feature values is mapped to {+1, -1} [5].

$$sign(x) = \begin{cases} 1, & \text{for } x \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{1.3}$$

$$y = sign(\boldsymbol{X} \cdot \boldsymbol{W} + b) = sign(\sum_{i=1}^{n} x_i w_i + b) \tag{1.4}$$

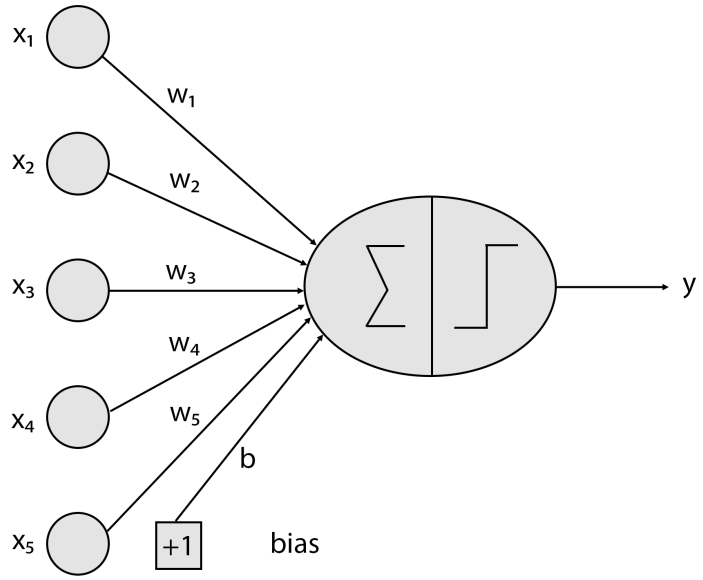

Figure 1.1: The figure illustrates the architecture of a perceptron with inputs $x_1, ..., x_5$, weights $w_1, ..., w_5$, a bias b and the resulting output y after the activation function.

1.1.2 Multilayer Perceptrons

Multilayer perceptrons (MLPs), also known as multilayer neural networks, consist of multiple layers of neurons. In addition to the perceptron, which has only one calculation layer, the output layer, MLPs use intermediate layers (hidden layers) to solve more complex tasks. The standard architecture of MLPs are known as *feed-forward networks*, where neurons of a layer are connected to neurons of the consecutive layer in forward direction [5]. Therefore, feed-forward networks do not incorporate feedback connections like loops to previous layers. If ANNs include feedback connections they are called *recurrent neural networks* [6].

1.1.3 Activation Functions

The choice of a proper activation function for a particular task and network architecture is crucial for balanced training and network convergence. They do

not only impact the convergence speed, but also the neural network's accuracy and computational efficiency. Different types and modifications of activation functions exist to satisfy different purposes.

The most commonly used function is the Rectified Linear Unit (ReLU) introduced by Nair and Hinton in [7], shown in figure 1.2 a). ReLU returns zero for any negative input, but returns any positive input unchanged:

$$f(x) = max\{0, x\} \tag{1.5}$$

Therefore, ReLU is a simple yet powerful non-linear function. Training with a ReLU activation function via gradient-based optimization has, however, one disadvantage: it is not able to learn from inputs that generate zero activations. General problems of training neural networks are exploding or vanishing gradients. The drawback of ReLUs are vanishing gradients, which is referred to as dying ReLU and can occur if many or all neurons enter an inactive state and output zero for any input [8].

One advancement is Leaky ReLU, which was invented by Maas, Hannun and Ng [9]. It is a more generalized modification of ReLU and uses a non-zero slope α for $x < 0$ to have gradients everywhere, as shown in equation 1.6. The standard value of α is 0.01. This small slope solves the problem of dying activations [6]:

$$f(x) = \begin{cases} x, & \text{for } x > 0 \\ \alpha x, & \text{otherwise.} \end{cases} \tag{1.6}$$

Before ReLU and its modifications were introduced, the sigmoid and hyperbolic tangent (tanh) activation functions were commonly used in neural networks:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \textit{ (sigmoid function)} \tag{1.7}$$

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \textit{ (hyperbolic tangent function)} \tag{1.8}$$

Both activation functions are closely related and have a similar shape, as shown in *c)* and *d)* of figure 1.2. The sigmoidal activations are in the range $[0, 1]$, which are vertically re-scaled to $[-1, 1]$ for the tanh activations [5]:

$$tanh(x) = 2\sigma(2x) - 1 \tag{1.9}$$

One backdraw of sigmoidal units is their difficulty to train, because of saturation. If x is very low or very large, the outputs saturate to the maximum or minimum, respectively. Therefore, piecewise-linear activation functions are preferentially used in hidden layers. Sigmoidal activation functions can be used as output units, representing a probabilistic output, although the tanh activation function performs better and is easier to train [6].

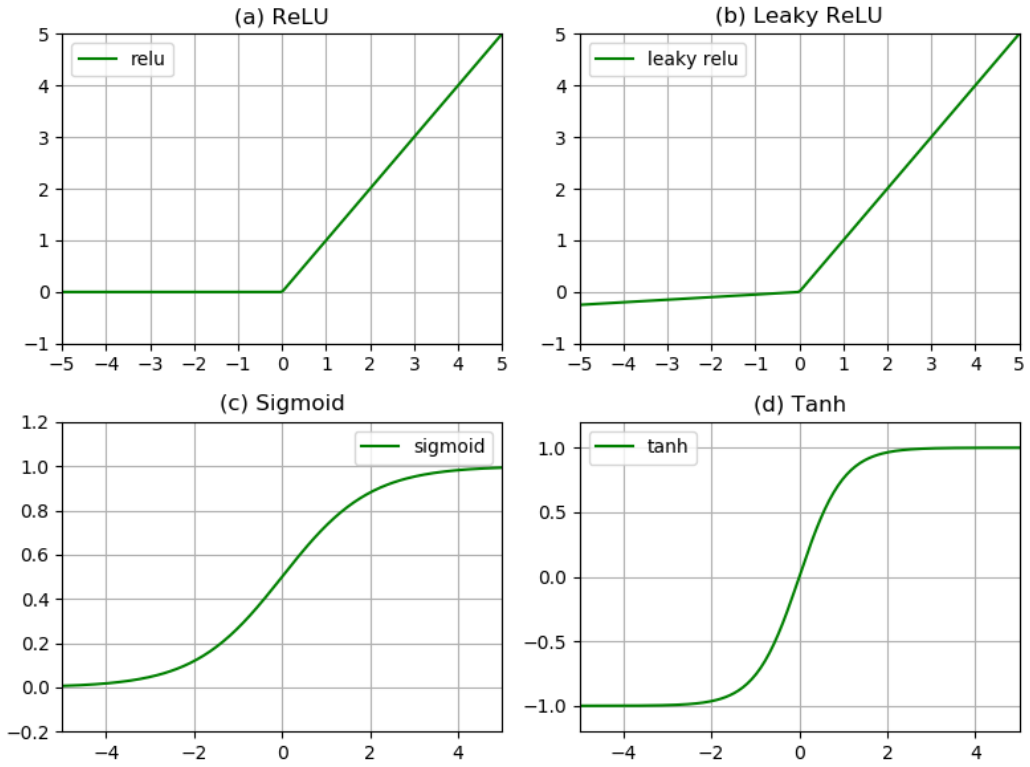

Figure 1.2: Various activation functions for artificial neural networks (ANNs).

1.2 Deep Learning

Deep learning refers to ANNs with a high amount of layers resulting in lots of parameters. Such networks are called deep neural networks (DNNs) and started to shine when computational power strongly increased and large datasets got

available. Since then, large models were introduced, which have millions of parameters and reach top accuracies in different domains, for example, speech or image processing [6].

1.2.1 Training a Deep Neural Network

DNNs distinguish between the training phase and the test (prediction) phase, in which a train and a test dataset are used, respectively. In a multi-layer feed-forward network the final output is calculated by the composed functions of each layer, for example $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ [6].

After calculating the network's output, the prediction error is calculated based on a specific loss function. In most deep learning projects, the network is optimized with some sort of gradient descent optimizer, which iteratively minimizes the total loss [6].

The gradients for the gradient-descent optimizers are calculated with the *back-propagation* algorithm, which consists of a forward and backward phase. During the forward phase, the network's output for the training sample is calculated using its current weights. In the backward phase, the gradients of the loss function are calculated with respect to the weights and used to update the weights for the next training iteration. The backpropagation algorithm uses the chain rule, which multiplies the partial derivatives of each node along a path to the output and sums up all such paths of the network. The chain rule's result is the derivative of the output with respect to the weights [5].

Therefore, choosing an appropriate network architecture, loss function and optimizer for a given task is crucial for a proper training of DNNs [6].

1.2.2 Loss Function

The loss function is one of the most important decisions to consider when designing ANNs. The result of the loss function states how much a learned and a target value differ.

A suitable choice for the loss function depends on the specific application and the activation function. There are various different loss functions, but some are

preferably used in combination with specific activation functions, for example, linear activation functions are commonly used with squared loss [5].

In the field of ANNs, it is common to learn conditional distributions by training with maximum likelihood. Therefore, the loss function is simply given by the negative log likelihood (NLL), which is stated in equation 1.10. The loss function for a specific parametric model p_{model} is derived from the maximum likelihood and can vary between different models. When it comes to designing loss functions the total loss function used to train ANNs often consists of an elementary loss function term for target-performance comparison and a second term for regularization [6].

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{x,y\sim\hat{p}_{data}} \log\, p_{model}(\boldsymbol{y}|\boldsymbol{x}) \tag{1.10}$$

1.2.2.1 Binary Cross Entropy Loss

A common loss function for logistic regression is the cross-entropy (CE) loss. Considering a classification task, a given observation x and an outcome y, the CE loss is some measure of how good the classifier's prediction $\hat{y}$ is, compared to y. If there are two different classes, it is referred to as binary classification and the corresponding loss function is the binary cross-entropy (BCE) loss, which is based on the Bernoulli distribution. This distribution has a binary output: 1 with probability p and 0 with probability $1 - p$.

The probability output of the classifier for one observation x can be expressed as:

$$p(y|x) = \hat{y}^y(1-\hat{y})^{(1-y)} \text{ where } y = [0,1],\ \hat{y} = [0,1] \tag{1.11}$$

The BCE loss is obtained by taking the logarithm of equation 1.11 for computational simplification and changing the sign to negative to switch from a maximization to a minimization problem [10]:

$$\mathcal{L}_{BCE}(\hat{y}, y) = -\Big[y \log \hat{y} + (1-y) \log (1-\hat{y})\Big] \tag{1.12}$$

1.2.2.2 L1 Loss

The L1 loss calculates the mean absolute error (MAE) between all elements of input x and target y [11]:

$$\mathcal{L}_{L1}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - y_i \right| \tag{1.13}$$

1.2.2.3 Kullback-Leibler Divergence

Considering two distributions p(x) and q(x), it is possible to express their similarity with the Kullback-Leibler Divergence (KLD). If p(x) is some unknown distribution and q(x) is used to approximate p(x), then the result of KLD states how good the approximation is [12]:

$$\begin{aligned} D_{KL}(p||q) &= -\int p(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x} - (-\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x}) \\ &= -\int p(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}) \end{aligned} \tag{1.14}$$

Important properties of KLD are on the one hand its non-negativity and on the other hand its result being greater than zero, except if p(x) = q(x). There is one aspect to be aware of: KLD can not be seen as a measure of distance, because it is not a symmetrical metric $D_{KL}(p||q) \neq D_{KL}(q||p)$ [12].

1.2.2.4 Perceptual Loss

Conventional CNNs (section 1.2.5) calculate a per-pixel loss between a ground truth image and the network's output. Johnson, Alahi and Fei-Fei show in [13] how to calculate perceptual and semantic distinctions between images. They define a feature reconstruction loss, shown in equation 1.15, which not only compares an output with a ground truth but calculates image differences of CNN layers of different network depths. It does not aim to minimize a per-pixel

loss but tries to have a similar feature representation. According to them, it is possible to compare high-level differences and preserve spatial structures. They use a pre-trained VGG-16 [14] network ϕ, which is pre-trained on the ImageNet dataset [15], as perceptual loss function. $\phi_j(x)$ indicates the feature map of the j-th convolutional layer of the network ϕ, which has a dimension of $C_j \times H_j \times W_j$. The ground truth is noted as y, while the output image is noted as $\hat{y}$.

$$\mathcal{L}^{\phi,j}_{Percept}(\hat{y}, y) = \frac{1}{C_j H_j W_j} ||\phi_j(\hat{y}) - \phi_j(y)||_2^2 \tag{1.15}$$

1.2.3 Parameter Optimization

The weights and biases of the ANNs are altogether referred to as parameters θ, which are meant to be optimized by a chosen optimizer. There are several different algorithms suitable for parameter optimization, which are based on the gradient descent (GD) algorithm, which is described in section 1.2.3.1.

Learning rates are hyperparameters that are difficult to set and have a large impact on the execution time of ANNs. In high-dimensional parameter spaces, the loss function can be sensitive to some directions but insensitive to others. Therefore, it can make sense to use individual adaptive learning rates for each parameter [6]. The adaptive learning rate algorithms AdaGrad, RMSProp and ADAM are described in sections 1.2.3.2, 1.2.3.3 and 1.2.3.4, respectively.

1.2.3.1 Gradient Descent

GD is an iterative optimization algorithm that is used to minimize the loss by moving small steps in the direction of the negative gradient. There exist different advanced variants of GD, but the most popular is stochastic gradient descent (SGD) which is based on the stochastic approximation method by Robbins in [16]. The mathematical requirement to use GD is that the loss function is differentiable w.r.t its parameters [17].

The SGD algorithm updates the parameters by moving a small step in the direction of the steepest descent, which is represented by the negative gradient

of the loss function $\mathcal{L}(\boldsymbol{\theta}^{(\tau)})$ with respect to the parameters. The step size is controlled by the learning rate η, which can be only a positive number. The parameter update of the SGD algorithm is specified by [12]:

$$\boldsymbol{\theta}^{(\tau+1)} = \boldsymbol{\theta}^{(\tau)} - \eta \nabla \mathcal{L}_i(\boldsymbol{\theta}^{(\tau)}), \qquad \eta > 0 \tag{1.16}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \tag{1.17}$$

SGD updates the parameters on a single data sample i while batch GD, which is a closely related GD method, takes all data samples at once into account. If the number of used data samples is in the range between one and all samples, then it is called minibatch gradient descent [6].

1.2.3.2 AdaGrad

AdaGrad is an adaptive learning algorithm described in [18]. The advantages according to the paper are sparse solutions, naturally incorporated regularization and better performance than non-adaptive methods. It is designed to converge fast on convex functions.

AdaGrad adapts the learning rates per parameter. It keeps track of the history of squared gradients. It updates the learning rates by scaling them inversely proportional to the square root of accumulated previous gradients. AdaGrad reduces the learning rates of the parameters appropriate to the value of their partial derivatives of the loss function. Consequently, learning rates of parameters with high partial derivatives are stronger reduced than the ones with lower partial derivatives [6].

1.2.3.3 RMSProp

RMSProp is another algorithm for adaptive learning rates and is described by Hinton in [19]. It is similar to AdaGrad with an adaption of how the history of gradients is computed. In contrast to AdaGrad, which uses accumulation

of the gradients, RMSProp applies exponentially decaying averaging and does not take older gradients of the history into account. Therefore, it can converge fast after locating a convex bowl. It is an effective algorithm and used in deep learning projects [6].

1.2.3.4 Adaptive Moment Estimation

Adaptive Moment Estimation (ADAM) is used for stochastic optimization and is gradient-based like SGD. ADAM is a first-order optimizer and its advantages are memory and computational performance. It is also well suited for networks with a large set of parameters. ADAM calculates parameter specific adaptive learning rates based on the gradients' first and second moments. The algorithm combines the advantages of RMSProp and AdaGrad [17]. ADAM is a robust algorithm concerning hyperparameter selection. [6].

1.2.4 Issues of Training

Training ANNs can be a complex task and some issues can arise. The source of the issues can be versatile, ranging from using activation functions that encourage vanishing or exploding gradients to training too complex models with too little data. The next sections describe some common issues and describe solutions to get them under control.

1.2.4.1 Overfitting and Underfitting

Considering an ANN trained and tested on training data and test data, respectively, the performance between both sets will be different. The term overfitting refers to the circumstance, that it is not ensured that a network will achieve good results on the test set, although it is perfectly fit to the training data. This difference in performance increases if a model with high capacity is trained on a small dataset [5]. The overfitting effect is illustrated in the right model of figure 1.3.

Another challenge in training neural networks is underfitting, which describes the situation of an ANN not achieving a low enough training error on the

dataset [6]. The underfitting effect is illustrated in the left model of figure 1.3.

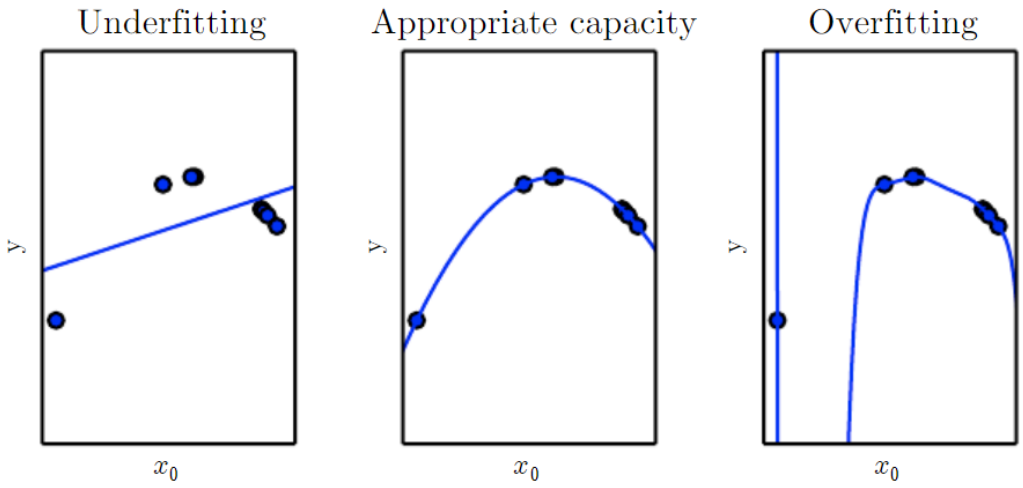

Figure 1.3: All three models are fit to an example dataset. The left model is clearly underfitting the training data and has a high error, while the right model has a low error and perfectly fits the training data, but will not be able to perform well on unseen data, due to overfitting. The model in the middle fits the training data well and will also perform good on unseen data (from [6]).

Considering a simple model, the gap between training and generalization error is small. However, it is desired to have a low training error and not just take a simple model to have a low gap between these errors. Usually, with increasing model capacity the training and generalization error decrease until the optimal capacity is reached, which is indicated as underfitting zone. If the model gets too complex, the generalization error rises again, which leads also to a higher gap between training and generalization error, which describes the overfitting zone [6]. Figure 1.4 illustrates the tradeoff between error and capacity.

1.2.4.1.1 Regularization

One technique to prevent the network from overfitting is regularization. The network's complexity can be regulated by adding a regularization term to

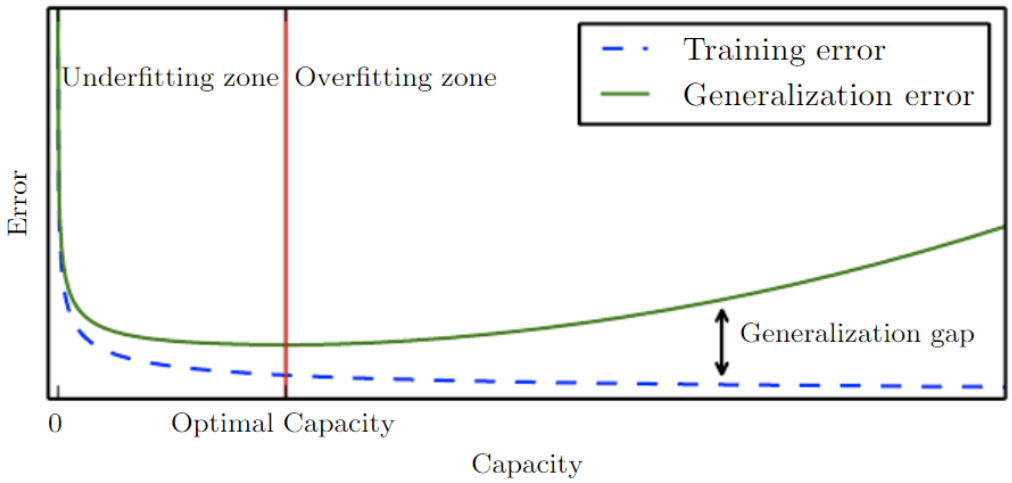

Figure 1.4: The left side of the red line represents the underfitting zone with both errors beeing high. The right side illustrated the overfitting zone, where the model capacity is increased but the generalization gap also increases. The optimal model capacity is described by the tradeoff between error and capacity when the training error and generalization gap are low and the generalization error is at the lowest level just before increasing again (from [6]).

the loss function. One such regularizer is the quadratic, also known as weight decay:

$$\widetilde{\mathcal{L}}(w) = \mathcal{L}(w) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{1.18}$$

where λ controls the amount of regularization [12].

Minimizing $\widetilde{\mathcal{L}}(w)$ results in a tradeoff between fitting the training data and ensuring to keep the weights low. Regularization is only meant to affect the generalization error, not the training loss [6].

There are various types of regularization, which serve different tasks. A commonly used regularization method is *early stopping* and is used to stop the training process if the error on a held-out dataset starts to rise again. The effect of this method is, that it regularizes the possible parameter space to be in a close region to these of the initialization stage [5].

1.2.4.1.2 Data Augmentation

In image segmentation tasks, a high amount of images is beneficial to successfully train high-performance segmentation networks, which is rarely the case in the domain of medical image segmentation.

Data augmentation solves this issue by applying different techniques to increase the amount of data. With the help of augmentation techniques datasets are filled with fake/synthetic data [6].

Data augmentation is a convenient way to lower the generalization error and can be seen to be associated with preprocessing. The techniques related to data augmentation are, for example, random translations and rotations of images and are only applied to the training data [6]. It is important to mention, that most of these (simpler) techniques can be performed during the training process, due to not being computationally expensive. Further techniques can be reflection, patch extraction or a computational extensive one like PCA transformation. Despite of data augmentation reducing issues of overfitting, the applied methods need to match the purpose of the ANNs. For example, it will not make sense to additionally train with mirrored images on the MNIST handwritten digits dataset [5]. Learning to detect a mirrored nine for example does not make any sense in common scenarios.

1.2.4.2 Vanishing and Exploding Gradients

As mentioned in section 1.2.1, the backpropagation is done using the chain rule. The drawback of this operation is a lack of stability of the gradient updates if the network consists of many layers. Therefore, gradients of the earlier layers can be extremely small, which is known as vanishing gradients, or extremely large, which is known as exploding gradients. The effect of vanishing or exploding gradients is common in DNNs. The origin of this issue is the chain-like product calculation of derivatives. On the one hand, if the gradients are mostly lower than 1 the products fall off exponentially along the chain and on the other hand, if they are mostly higher than one, then the products rise exponentially. There are activation functions that decrease such effects like ReLU with a gradient of one for positive values, but also some which encourage such issues, for example

a sigmoid activation function with a gradient lower than 0.25 and therefore prone to vanishing gradients [5].

1.2.4.3 Computational Issues

Training ANNs on image data is a highly expensive computational task and can take from many hours to weeks, due to the reason of datasets consisting of up to millions of images. The runtime of these training processes can be decreased with technological improvements, like the usage of Graphical Processor Units (GPUs). Therefore, it is useful to work with machine learning frameworks that have GPU support [5].

1.2.5 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a specialized type of ANNs suitable for processing grid-like data, like images. For this reason, CNNs are widely used in image processing [6]. They are based on biological insights of image processing in the visual cortex of cats by Wiesel and Hubel in [20], who figured out elements of the brain's processing on visual perception.

The problem with fully-connected neural networks and input data like images is, that the number of weights can increase fast, which can lead to a high amount of memory requirement. MLPs with a fully-connected first hidden layer with several hundreds of neurons would need to learn already millions of weights, considering a typical input image with ten thousands of pixels. A large number of parameters results in a more complex model, which increases the size of the needed training set [21]. In comparison to MLPs, in which weights are not shared and only used once for an output, CNNs share weights across multiple inputs. This means that the same element of a kernel is shared across all pixels of the image, concerning image processing. Learning fewer parameters results in a lower memory requirement and increased efficiency [6].

CNNs consist of different types of layers: convolution, pooling and fully-connected layers. These layers can be repeatedly stacked on each other like in the LeNet-5 architecture described in [21] and shown in figure 1.5.

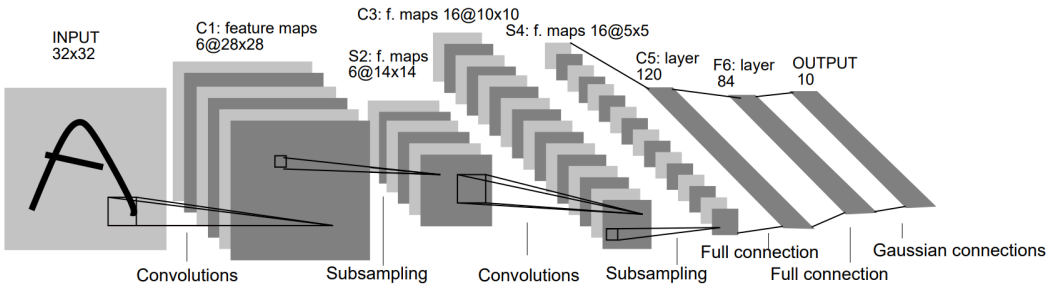

Figure 1.5: Example of a Convolutional Neural Network (CNN) representing the architecture of LeNet-5, which consists of convolution, pooling (sub-sampling) and fully-connected layers (from [21]).

1.2.5.1 Convolution Layer

The convolutional layer performs a convolution operation, which convolves an input image I with a kernel K, which is, for example, in the two-dimensional case described in the following formula [6]:

$$F(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{1.19}$$

Convolving an input with a filter kernel results in a so-called feature map. The convolution layer usually consists of several filter kernels each resulting in an individual feature map. This enables to extract multiple features at the same locations. A local receptive field describes the area of the input, which is used for the extraction of visual features like edges and corners [21].

There are two important parameters of the convolution layer, which have an impact on the feature map size: padding and stride. Padding refers to how the kernel is moved across the border of the image. It is differentiated between valid, half and full padding. *Valid padding* means the kernel is moved exactly across the image, and therefore, the dimension is reduced compared to the input. Considering a filter of size F_s, *half-padding* means, that $(F_s - 1)/2$ pixels are added at the border of the image, which guarantees, that the spatial dimension of the image is not changed. *Full padding* means, that the image is padded with $F_s - 1$ pixels, which leads to an increase of the output dimension compared to the input. The stride parameter enables to decrease the output dimension of

the convolution. A stride s of one means, that the kernel is moved by one pixel and the spatial dimension is unchanged. The spatial dimension is reduced by approximately s for $s > 1$. Strides larger than two are rarely used [5].

An example of a convolution operation of an RGB input image (32x32x3) with multiple kernels, with padding and stride of one, can be seen in 1.6. The filter kernel (5x5x3) is slid over the whole image at each location the convolution is performed. The result of the convolution of the input image with one kernel results in a feature map of 32x32x1. In this example, ten filter kernels are used, which results in an output volume of 32x32x10 consisting of ten feature maps [22].

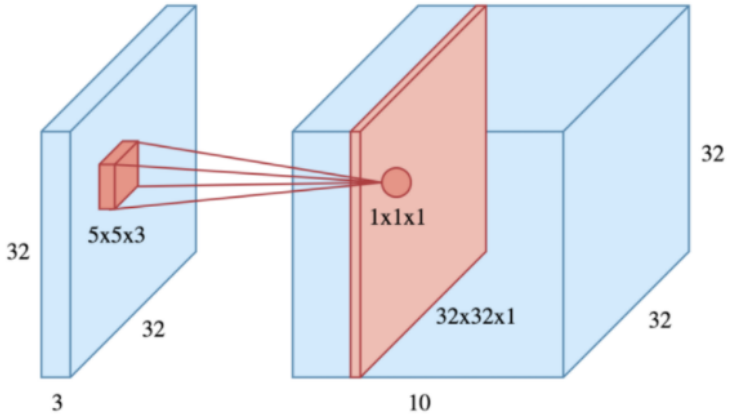

Figure 1.6: Convolving an RGB input image of 32x32x3 with a kernel of 5x5x3 with padding and stride of one results in a feature map of 32x32x1. Considering ten filter kernels, the convolution layer results in a volume of 32x32x10, representing the feature maps (from [22]).

After the convolution layer a nonlinear activation function, for example a ReLU activation function, is applied to the output [6].

1.2.5.2 Pooling Layer

The pooling or sub-sampling layer performs a local operation in a defined window, which reduces the dimensions of feature maps. Furthermore, the

sensitivity to feature shifts and distortions is lowered [21]. The most popular pooling operations are average-pooling (used, for example, in the LeNet-5 architecture), which calculates the average of the window's elements and max-pooling, which computes the maximum value of the window's elements. Similar to the convolution layer a stride parameter defines the step size of the sliding window [5].

1.2.5.3 Fully Connected Layer

The fully connected layer is equal to those in conventional feed-forward networks and connects all neurons of a layer with those of a consecutive layer. CNNs can consist of multiple fully connected layers [5]. The last fully connected layer is then connected to the output layer. The structure of the output layer depends on the task.

1.2.5.4 Transposed Convolution Layer

Convolution layers can be used in a way that leads to a decrease in spatial dimensionality. The opposite effect can be accomplished by either using upsampling layers or transposed convolution layers. Both layers lead to an increase in dimensionality. The difference between these two layers is that upsampling is done by interpolation without learned parameters and transposed convolution learns parameters during the training process. It is important to note that transposed convolutions are not an inverse transformation of convolutions on a value basis, but in dimensionality. Transposed convolution layers swap the forward and backward pass in comparison to common convolutional layers [23].

1.2.6 Autoencoder

The basic idea of an autoencoder is to reconstruct the given input. An autoencoder consists of an *encoder* function, a *latent space* and a *decoder* function, as shown in figure 1.7. While the *encoder* function $h = f(x)$ learns to represent the input data, the decoder function $x' = g(f(x))$ tries to reconstruct the input

from the encoder's output. The autoencoder network is regularized to prevent it from just copying the input to the output. This is achieved by constraining h to be of a smaller dimension than the input x. When training an autoencoder the *latent space* mapping is learned, which captures the most salient features in the input data. This bottleneck layer captures a compressed representation of the more complex input data and is represented by the nodes in the green rectangle of figure 1.7. Training an autoencoder is performed by minimizing a loss function $\mathcal{L}(\mathbf{x}, g(f(\mathbf{x}))$ [6].

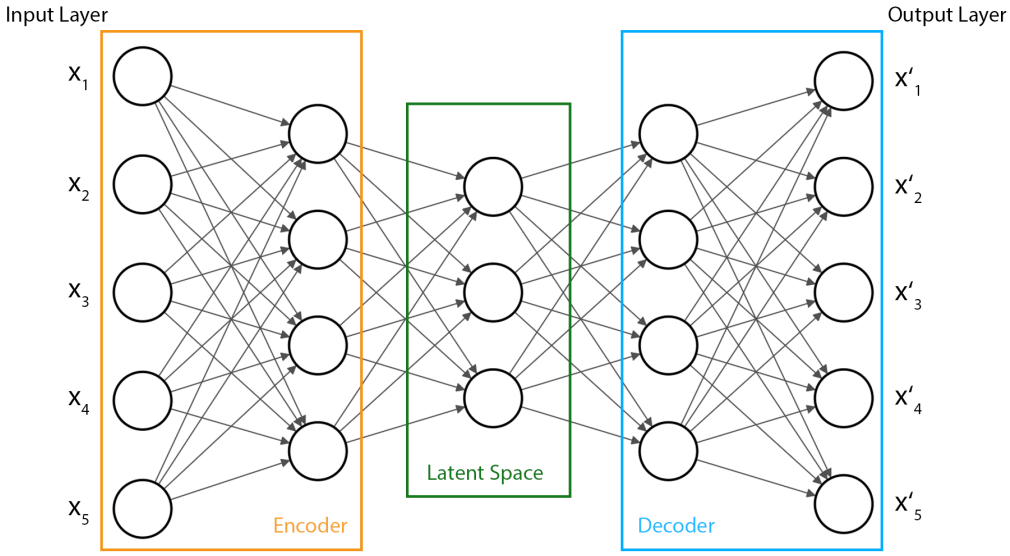

Figure 1.7: Basic scheme of an autoencoder with a single hidden layer.

Due to these optimization steps, the weights of the encoder and decoder structures are changed. The final reconstruction resembles the input data, however, this reconstruction of the compressed latent space is lossy. Autoencoders are fundamental networks of unsupervised learning tasks. Some example applications are dimensionality reduction, outlier detection or de-noising [5].

1.2.7 Variational Autoencoder

The basic structure of variational autencoders (VAEs) resembles that of classic autoencoders: both consist of an encoder and decoder structure. The main

difference between both networks is located in the latent space. This section is based on the works of Kingma and Welling in [24] and [25]. The VAEs regularize the latent space in a probabilistic way. The latent space of VAE consists of one layer representing standard deviations and a second one representing the means.

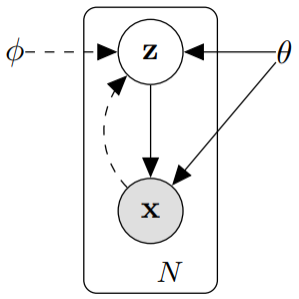

Figure 1.8: Graphical model of the variational autoencoder's latent variable $\mathbf{z}$ and observed variable $\mathbf{x}$ representing the data. The generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})$ is illustrated by solid lines, while the dashed lines represent the posterior approximation $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ with variational inference (from [24]).

In the probabilistic graphical model in figure 1.8, $\mathbf{x}$ identifies the observed variable, which represents the input data and $\mathbf{z}$ denotes the latent variable, which is assumed to be drawn from a Gaussian prior distribution with zero mean and identity covariance matrix:

$$p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I}) \tag{1.20}$$

where θ denotes the generative model parameters. Subsequently, the data $\mathbf{x}$ is sampled from the conditional likelihood distribution $p_\theta(\boldsymbol{x}|\mathbf{z})$. Considering the VAE's scheme in figure 1.9, $p_\theta(\boldsymbol{x}|\mathbf{z})$ is also referred to as the *probabilistic encoder*, whereas $p_\theta(\boldsymbol{z}|\mathbf{x})$ is the posterior distribution and referred to as the *probabilistic decoder*. The likelihood function models the distribution of $\mathbf{z}$ given a data sample $\mathbf{x}$, while the posterior models the distribution of the potentially related data sample $\mathbf{x}$ given an encoded latent representation $\mathbf{z}$.

The relations of the mentioned distributions are explained by the Bayes theorem:

$$p_\theta(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})}{p_\theta(\boldsymbol{x})} = \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})}{\int_z p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})d\boldsymbol{z}} \tag{1.21}$$

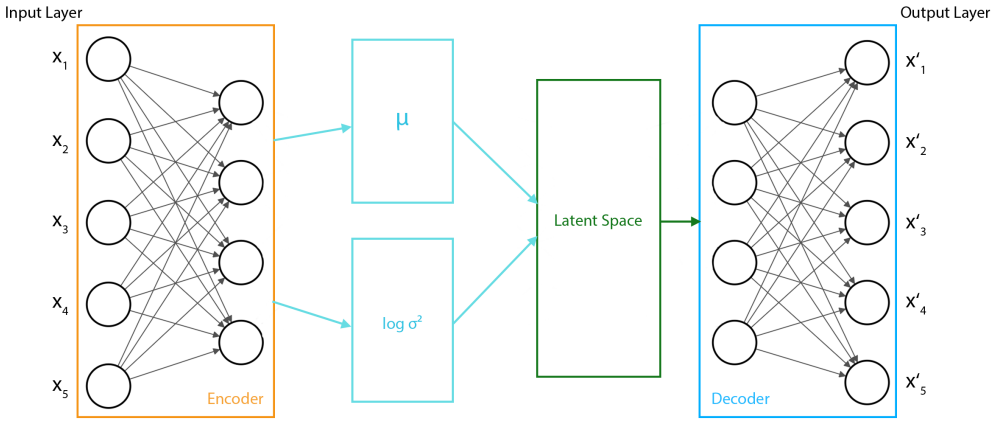

Figure 1.9: Basic scheme of a variational autoencoder. The output of the encoder acts as input of a μ and $log\sigma^2$ layer. The latent space samples latent variables z as in equation 1.23.

Theoretically, if $p_\theta(\mathbf{z})$ and $p_\theta(\boldsymbol{x}|\mathbf{z})$ are known, then $p_\theta(\boldsymbol{z}|\mathbf{x})$ could be calculated with the use of the Bayes theorem. Due to the integral of the marginal likelihood $p_\theta(\boldsymbol{x})$ becoming intractable in high dimensions, the Expectation-Maximization (EM) algorithm cannot be used for parameter estimation and an approximation of the posterior distribution $p_\theta(\boldsymbol{z}|\mathbf{x})$ by $q_\phi(\boldsymbol{z}|\mathbf{x})$ is needed. This approximation of the posterior is achieved with variational inference. $\boldsymbol{\phi}$ denotes the variational parameters in $q_\phi(\boldsymbol{z}|\mathbf{x})$.

Optimizing VAEs means optimizing the *evidence lower bound* (ELBO), which is typically derived by Jensen's inequality. Kingma and Welling show a reparameterization trick on the variational lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)})$, which results in an estimator, which can be optimized using standard stochastic gradient methods. The resulting stochastic gradient variational bayes (SGVB) estimator for data sample $x^{(i)}$ is:

$$\begin{gathered}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) = -D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p_\theta(\boldsymbol{z})) + \frac{1}{L}\sum_{l=1}^{L}(log\ p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)})) \\ \text{where } \boldsymbol{z}^{(i,l)} = g_\phi(\boldsymbol{\epsilon}^{(i,l)}, \boldsymbol{x}^{(i,l)}) \text{ and } \boldsymbol{\epsilon}^{(l)} = p(\boldsymbol{\epsilon})\end{gathered} \tag{1.22}$$

The first term on the right side is the Kullback-Leibler divergence (KLD), which acts as regularizer and constrains the approximate posterior to be in close range to the prior $p_\theta(\mathbf{z})$, while the second term can be seen as reconstruction

error. Considering the KLD term, it is assumed, that both distributions, the posterior and prior are Gaussian. This leads to the possibility to analytically calculate the KLD term, as shown in formula 1.24. The function $g_\phi(\cdot)$ performs a mapping of the data sample $\boldsymbol{x}$ and random noise vector $\boldsymbol{\epsilon}^{(l)}$ in a way, that $\boldsymbol{z}^{(i,l)}$ can be sampled from the approximate posterior distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)})$.

There is a problem with the described random sampling because the random variable z is not differentiable. Therefore, the required back-propagation is not possible [5]. The before mentioned reparameterization trick is the solution to this shortcoming. Kingma and Welling propose a valid reparameterization of

$$z = \mu + \sigma \cdot \epsilon, \quad \text{for } z \sim p(z|x) = \mathcal{N}(\mu, \sigma^2). \tag{1.23}$$

This reparameterization changes z to a function of x and ϕ which is now differentiable and deterministic, because the randomness is added by a separate random variable ϵ. This enables the required back-propagation of the loss. A graphical model of the reparameterization trick is shown in figure 1.10.

This leads to the following loss function for VAEs:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) = \underbrace{\frac{1}{2}\sum_{j=1}^{J}(1 + log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2) - (\sigma_j^{(i)})^2)}_{KLD} + \frac{1}{L}\sum_{l=1}^{L}(log\ p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}))$$
$$\text{where } \boldsymbol{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \cdot \boldsymbol{\epsilon}^{(i)} \text{ and } \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \boldsymbol{I}) \tag{1.24}$$

1.3 Image Segmentation

Image Segmentation is a complex task and describes the procedure of splitting an image into regions of equal properties. Such homogeneous regions can be of the same brightness, contrast, color or gray-level. Medical image segmentation is used, for example, to analyze anatomical structures or locate regions of interest like tumors. It is difficult to automatically segment medical images due to the complexity of such images and additional artifacts. Some examples of common artifacts are motion artifacts, which are often related to scans of the thorax

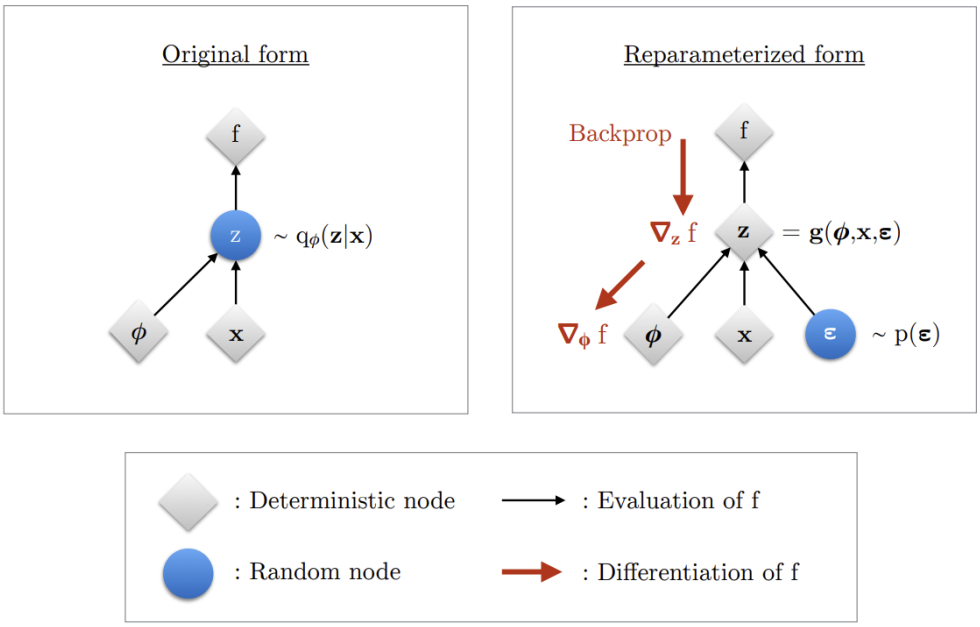

Figure 1.10: In the original form a back-propagation is not possible due to the not differentiable random variable z. The proposed reparameterization trick by Kingma and Welling introduces a new random variable epsilon. Z is now a function of ϕ and x and becomes differentiable and deterministic due to the shift of the random source (from [25]).

(due to breathing), noise artifacts, ring artifacts or intensity inhomogeneity [26].

The following methods are related to medical image segmentation of computed tomography (CT) or magnetic resonance imaging (MRI) scans and adapted from [26]. They can be roughly divided into methods based on gray level features, texture features as well as model, atlas and neural networks-based segmentations.

1.3.1 Gray Level Features based Segmentation

This set of methods segment images based on gray level information. They can be further divided into segmentation based on amplitude, edge-based segmentation and region-based segmentation.

1.3.1.1 Amplitude Segmentation

This methods are based on histogram information of images. A common method is thresholding, either based on histogram features or gray levels. This method is quite simple in the case of an image consisting of a single object and a background, which is represented by a bi-modal histogram. This results in a segmentation of the image into object and background. For a proper segmentation the threshold value needs to be chosen carefully. Multi-object segmentation is also possible if the objects differ in gray levels.

1.3.1.2 Edge-based Segmentation

Edge-based segmentation methods are based on the idea that edges represent boundaries between different objects. Therefore, these methods detect edges first, then perform a thresholding on the found edges and finally segmenting the images with the obtained boundaries. Problems arise with noise and weak edges, which can impact the final segmentation result.

1.3.1.3 Region-based Segmentation

These methods analyze the images for homogeneous regions and create clusters out of the corresponding pixels. The property used when deciding about similarities is the gray level. Three different types of region-growing algorithms exist: region merging, region splitting and split and merge. Issues and limitations of region growing algorithms can be under and over segmentation. One solution to this behaviour is to combine edge-based segmentation with region-based segmentation, which can improve the accuracy of an overall segmentation.

1.3.2 Methods based on Texture Features

This group of methods performs segmentations based on different textures in an image. These textures can differ in tone and structure, can be fine or coarse. There exist three different approaches: statistical approaches, syntactic or structural approaches and spectral approaches. The statistical approaches are useful for random and complex textures, however, the syntactic or structural approaches can lead to better segmentation results. Spectral approaches can be robust, however, are often not very efficient.

1.3.3 Model-based Segmentation

This method is based on the idea that structures like organs are similar in shape and geometry across different patients, although with some variations. This variations can be probabilistically described. Model-based segmentation methods are active shape and appearance models, deformable models and level-set-based models.

1.3.4 Atlas-based Segmentation

This segmentation method describes all properties regarding a targeted anatomy as an atlas or lookup table (LUT). Properties like the shape and size of organs, bones and soft tissues are encoded in such an atlas. The limitations of atlas-based segmentations are segmentations of complex structures with a variability of these properties.

1.3.5 Neural Networks-based Segmentation

Although, some of the methods mentioned before can produce reasonable segmentation results, image segmentation is mostly done with neural networks nowadays, because of their fully-automatic nature, requiring also no (manual) parameter adaptions. Deep learning had a big impact on the performance and results of such automatic methods. Image segmentation ANNs can be split into semantic segmentation and instance segmentation. Semantic segmentation

tries to separate a foreground object from a background, for example, a horse from the meadow. Instance segmentation, on the other hand, tries to find all instances of an object and individually segment them.

In semantic segmentation, there exist several well-known networks in the community. Some notable works are, for example, CNN-based networks like FCN [27], DeepLab [28], AlexNet [29] or encoder decoder structured networks like U-Net [30]. One of the well-known instance segmentation networks is Mask R-CNN [31].

1.4 PyTorch

PyTorch is an open source machine learning framework for Python, which also integrates with libraries like NumPy and SciPy. Furthermore, high computational tasks can be accelerated due to its CUDA integration. The following sections explain the basic structure and concepts of PyTorch and are adapted from the PyTorch documentation [32].

1.4.1 Torchvision

A major package of PyTorch is called Torchvision, which is a collection of datasets, models and useful image transformations. Amongst others, it includes the VGG model already mentioned in section 1.2.2.4.

1.4.2 Torch

This package contains the fundamental data structures called tensors for storing data, tensor operations and much more utility functionality. It provides the fundamental neural network layers, for example convolutional layer and pooling layer, which are located in the module `torch.nn`. There is a specific `torch.cuda` package, which enables to shift computations onto a GPU.

1.4.3 Tensors

Tensors are the way PyTorch stores and transfers data between operations. They are multi-dimensional matrices that can be either stored as CPU or GPU tensors. PyTorch offers the most common tensor types with different sizes, for example, integer, floating point and boolean. The tensor types range from 8 to 128 bit.

1.4.4 Training a neural network in PyTorch

Training a basic neural network in PyTorch needs at first a defined and read in dataset. Therefore, PyTorch provides a package `TORCH.UTILS.DATA`, in which a `Dataset` and `Dataloader` class are located. These classes handle settings, like batch size, data shuffling and more. Furthermore, a network needs to be defined. A network can be inherited and extended from the `Module` class in the package `TORCH.NN` or an already pre-defined one can be chosen. Additionally, an optimizer needs to be chosen. PyTorch provides some optimizers in the package `TORCH.OPTIM`. Finally, the training process needs to be implemented. A basic classifier example of the PyTorch Documentation is shown in figure 1.11. While looping over the batches of the dataloader the optimizer's gradients need to be reset to zero in each iteration. Subsequently, the input is passed to the network and a loss is calculated between the result and some ground truth labels. The final steps are back-propagating the loss with `loss.backward()` and updating the network's parameters with `optimizer.step()`.

1.5 CUDA

CUDA (Compute Unified Device Architecture) is a platform developed by Nvidia that enables software developers to use Nvidia GPUs for complex parallel computations [34]. Therefore, GPUs of Nvidia are used in the fields of deep learning to offer the opportunity to train large ANNs. PyTorch has integrated CUDA support with the `TORCH.CUDA` package [35], which enables easy usage of the GPU without the need of many instructions.

```
for epoch in range(2):  # loop over the dataset multiple times

    running_loss = 0.0
    for (i, data) in enumerate(trainloader, 0):

        # get the inputs; data is a list of [inputs, labels]

        (inputs, labels) = data

        # zero the parameter gradients

        optimizer.zero_grad()

        # forward + backward + optimize

        outputs = net(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
```

Figure 1.11: A PyTorch example for the training of a basic classifier (reproduced from [33]).

Bibliography

[1] J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek, "Medical deep learning – a systematic meta-review," *Computer Methods and Programs in Biomedicine*, p. 106 874, 2022, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.106874. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722002565 (cit. on p. 1).

[2] W. S. McCulloch and W. Pitts, *A Logical Calculus of the Ideas Immanent in Nervous Activity*. 1943, pp. 115–133. DOI: 10.1007/BF02478259 (cit. on p. 2).

[3] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *European journal of gastroenterology & hepatology*, vol. 19, pp. 1046–54, Jan. 2008. DOI: 10.1097/MEG.0b013e3282f198a0 (cit. on p. 2).

[4] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, ser. Report: Cornell Aeronautical Laboratory. Cornell Aeronautical Laboratory, 1957. [Online]. Available: https://books.google.de/books?id=P_XGPgAACAAJ (cit. on p. 2).

[5] C. C. Aggarwal, *Neural Networks and Deep Learning, A Textbook*. Springer, 2018, p. 497, ISBN: 978-3-319-94462-3. DOI: 10.1007/978-3-319-94463-0 (cit. on pp. 2–4, 6, 7, 11, 13–15, 17–19, 22).

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org (cit. on pp. 3–7, 9–17, 19).

[7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of International Conference on Machine Learning (ICML)*, 2010, pp. 807–814 (cit. on p. 4).

[8] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, *Dying relu and initialization: Theory and numerical examples*, 2019. arXiv: 1903.06733 (cit. on p. 4).

[9] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013 (cit. on p. 4).

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition)*. 2019 (cit. on p. 7).

[11] *L1loss - pytorch 1.6.0 documentation*, `https://pytorch.org/docs/stable/generated/torch.nn.L1Loss.html`, [Online; accessed 16-August-2021] (cit. on p. 8).

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738 (cit. on pp. 8, 10, 13).

[13] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. arXiv: `1603.08155`. [Online]. Available: `http://arxiv.org/abs/1603.08155` (cit. on p. 8).

[14] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: `1409.1556` (cit. on p. 9).

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. DOI: `10.1007/s11263-015-0816-y` (cit. on p. 9).

[16] H. E. Robbins, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 2007 (cit. on p. 9).

[17] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. arXiv: `1412.6980 [cs.LG]` (cit. on pp. 9, 11).

[18] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011 (cit. on p. 10).

[19] T. Tieleman and G. Hinton, *Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning*, Technical report, 2012 (cit. on p. 10).

[20] D. H. Hubel and T. N. Wiesel, *Receptive fields of single neurons in the cat's striate cortex*. 1959, pp. 574–591 (cit. on p. 15).

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradientbased learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324 (cit. on pp. 15, 16, 18).

[22] A. Dertat, *Applied deep learning - part 4: Convolutional neural networks*, `https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2`, [Online; accessed 16-August-2021] (cit. on p. 17).

[23] V. Dumoulin and F. Visin, *A guide to convolution arithmetic for deep learning*, 2018. arXiv: `1603.07285 [stat.ML]` (cit. on p. 18).

[24] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2013. arXiv: `1312.6114` (cit. on p. 20).

[25] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. arXiv: `1906.02691`. [Online]. Available: `http://arxiv.org/abs/1906.02691` (cit. on pp. 20, 23).

[26] N. Sharma and L. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics / Association of Medical Physicists of India*, vol. 35, pp. 3–14, Apr. 2010. DOI: `10.4103/0971-6203.58777` (cit. on p. 23).

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: `http://arxiv.org/abs/1411.4038` (cit. on p. 26).

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. arXiv: `1606.00915`. [Online]. Available: `http://arxiv.org/abs/1606.00915` (cit. on p. 26).

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12, Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105 (cit. on p. 26).

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597 (cit. on p. 26).

[31] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. arXiv: 1703.06870. [Online]. Available: http://arxiv.org/abs/1703.06870 (cit. on p. 26).

[32] *Pytorch documentation*, https://pytorch.org/docs/stable/index.html, [Online; accessed 16-August-2021] (cit. on p. 26).

[33] *Pytorch 1.8.0 documentation - training a classifier*, https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html, [Online; accessed 16-August-2021] (cit. on p. 28).

[34] *Cuda high performance computing*, https://developer.nvidia.com/about-cuda, [Online; accessed 16-August-2021] (cit. on p. 27).

[35] *Pytorch 1.8.0 documentation - torch.cuda*, https://pytorch.org/docs/stable/cuda.html, [Online; accessed 16-August-2021] (cit. on p. 27).