Artificial Intelligence in Ophthalmology

Andrzej Grzybowski
Editor

Second Edition

Springer

Artificial Intelligence in Ophthalmology

Andrzej Grzybowski
Editor

Artificial Intelligence in Ophthalmology

Second Edition

Springer

Editor
Andrzej Grzybowski
Department of Ophthalmology
University of Warmia and Mazury
Olsztyn, Poland

Institute for Research in Ophthalmology
Foundation for Ophthalmology Development
Poznan, Poland

ISBN 978-3-031-83755-5 ISBN 978-3-031-83756-2 (eBook)
https://doi.org/10.1007/978-3-031-83756-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1 Artificial Intelligence in Ophthalmology: Promises, Hazards and Challenges

Andrzej Grzybowski

"If you do not get feedback, your confidence grows much faster than your accuracy"
Tetlock P., Gardner D. Superforcasting: The Art and Science of Prediction, Crown Publishing, 2016.

Abstract

The chapter presents the overview of the present achievements and applications of artificial intelligence in medicine, including the possible benefits for future medicine. It also presents the increase in published scientific studies using artificial intelligence in medicine in last decade.

The regulation frameworks for medical devices, including AI medical devices, in the USA and in the European Union is discussed. Moreover, the problem of access to reliable data is given with the explanation of transfer learning, generative adversarial networks, and continual machine learning. Some of the hazards, and challenges of the AI in ophthalmology are described, including privacy protection, testing AI algorithms on data sets that did not correspond real-world conditions, and its vulnerability to cybersecurity attacks. Finally, some aspects of cost-effectiveness of AI-based devices are presented.

Keywords

Medicine · Ophthalmology · Artificial intelligence · Deep learning · Machine learning · Artificial intelligence in ophthalmology · AI devices regulations · AI devices safety

The Promise of Artificial Intelligence

The term "artificial intelligence" (AI) was coined on August 31, 1955, when John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon submitted "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." [1, 2]. It was, however, Alan Turing who during a public lecture in London in 1947 mentioned computer intelligence, and in 1948 he introduced many of the central concepts of AI in a report entitled "Intelligent Machinery." [3]. Moreover, Turing proposed in 1950 the test, originally called the imitation game, and later known as the Turing

A. Grzybowski (✉)
Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland
e-mail: ae.grzybowski@gmail.com

A. Grzybowski
Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland

A. Grzybowski (ed.), *Artificial Intelligence in Ophthalmology*,
https://doi.org/10.1007/978-3-031-83756-2_1

test, as a way to confirm that the intelligent behavior of a machine was equivalent to that of a human. A human evaluator is asked to determine the nature of a partner (human or machine) based on a text-only conversation [1–3]. After decades of slow progress since the Turing test was proposed, AI has finally blossomed. Many new technologies and applications are available, and there is great enthusiasm about the promise of AI in health care. It holds the potential to improve patient and practitioner outcomes, reduce costs by preventing errors and unnecessary procedures, and provide population-wide health improvements. We have entered the fourth stage of the Industrial Revolution that began in the eighteenth century, and its defining feature may well be the use of AI technologies (Fig. 1.1). The results of an annual competition known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) provide interesting insights into recent developments in AI technology (Fig. 1.2). Over the years 2010–2016 there was a steady decrease in the error rates of the algorithms presented, and in 2017, 29 of the 38 competing teams had error rates lower than 5% (considered to be the human threshold). Thus in 10 years AI algorithms exceeded human performance in image recognition.

There are many promising applications for AI in health care, addressing a variety of aims and taking many different approaches (Table 1.1). For example, misdiagnoses constitute a huge, although poorly recognized, medical problem. A study published in 2014 estimated that diagnostic errors affect at least 5% of US adults (12 million people) per year [4]. More recently a systematic review and meta-analysis reported that the rate of diagnostic errors causing adverse events among hospitalized patients was 0.7% [5]. Furthermore, diagnostic error is the most important reason for malpractice litigation in the United States, accounting for 31% of malpractice lawsuits in 2017 [2]. The creation of AI programs to identify and analyze diagnostic errors could be an important step in addressing this problem [6].

Eric Topol has proposed that AI could help shift into "deep medicine," by allowing the physicians to devote more time to crucial relationships with their patients—an aspect of medicine that cannot be replaced by any AI technology [2]. It is also interesting to consider whether AI might enrich the doctor-patient relationship, enabling a shift from the present "shallow medicine" into "deep medicine," based on deep empathy and connection [2]. Success in building such relationships is very much related to the amount of time doctors can spare for patients and the extent of the personal contact they have with their patients. The average time of a clinic visit in the United States for an established patient is 7 min and for a new patient 12 min. In many Asian countries, clinic visits last as little as 2 min per patient [2]. Making this situation even worse, part of this time must be devoted to completing electronic health records, further limiting personal contact. A study published in

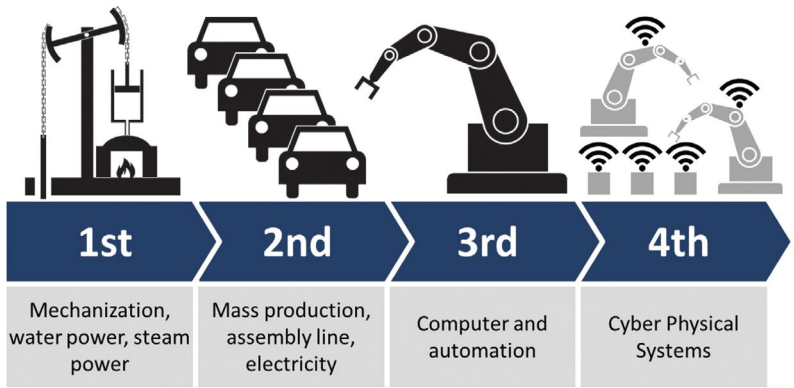

Fig. 1.1 The four main stages of the Industrial Revolution that began in the eighteenth century

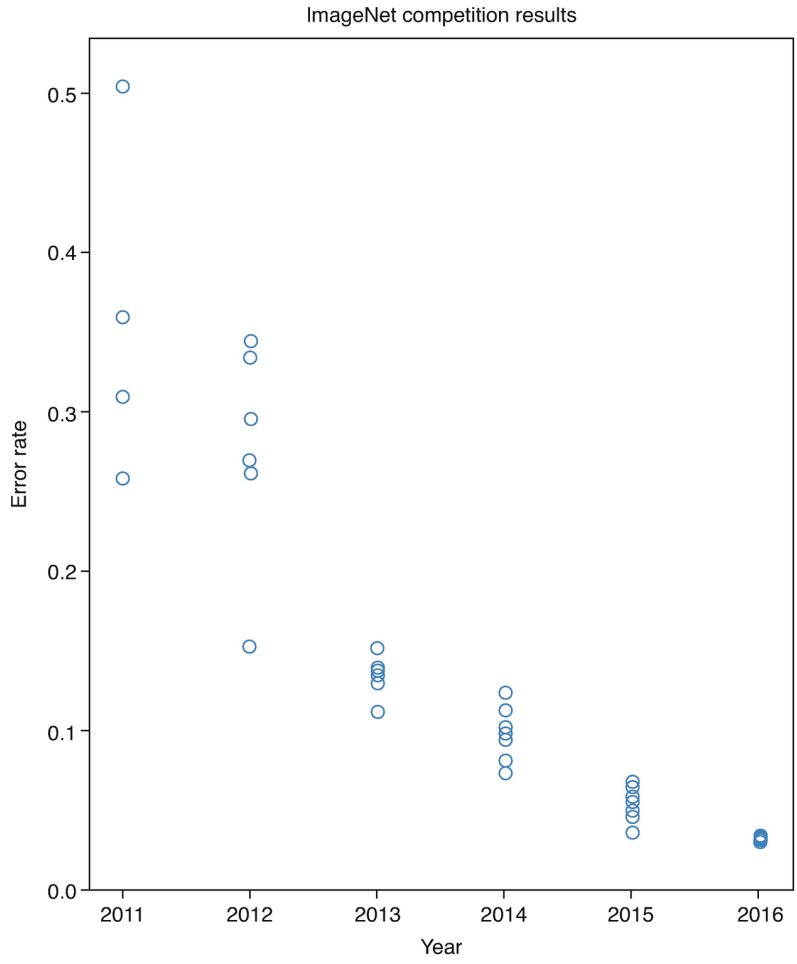

Fig. 1.2 Error-rate history on ImageNet

Table 1.1 Some ambitious expectations for AI in health care.

- outperform doctors,
- help to diagnose what is presently undiagnosable,
- help to treat what is presently untreatable,
- to recognize on images what is presently unrecognizable,
- predict the unpredictable,
- classify the unclassifiable,
- decrease the workflow inefficiencies,
- decrease hospital admissions and readmissions,
- increase medication adherence
- decrease patient harm
- decrease or eliminate misdiagnosis

Adapted from Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, New York 2019

2017 that asked patients to describe how they perceive their physician found that the most common negative responses were "rushed," "busy," and "hurried." [7]. These reactions are manifestations of "shallow medicine."

One of the arguments supporting the use of AI in medicine is that human cognitive capacity to effectively manage information is often exceeded by the quantity of data generated. Each year the world produces zettabytes of data (roughly, enough to fill a trillion smartphones) [2]. Moreover, unlike humans, who have bad days and emotions, and who get tired, with subsequent decreases in performance and accuracy, AI works 24/7 without vacations or complaints [2].

AI-based technologies employing deep-learning (DL) approaches have proven effective in supporting decisions in many medical specialties, including radiology, cardiology, oncology, dermatology, ophthalmology, and others. For example, AI/DL algorithms (also referred to as AI/DL models in the following text) have been shown to reduce waiting times, improve medication adherence, customize insulin dosages, and help interpret magnetic resonance images. The number of AI life-science papers listed in PubMed increased from 596 in 2010 to 12,422 in 2019 [8]. The number of papers on the use of AI in the field of ophthalmology also has increased dramatically (Figs. 1.3 and 1.4).

AI/DL algorithms have been used to detect diseases based on image analysis, with fundus photos and optical coherence tomography (OCT) scans analyzed for retinal diseases, chest radiographs assessed for lung diseases, and skin photos analyzed for skin disorders. Retinal photos have also been used to identify risk factors related to cardiovascular disorders, including blood pressure, smoking, and body mass index [9]. Using DL models trained on data from over 280,000 patients and validated on two independent data sets, Poplin et al. predicted cardiovascular risk factors not previously thought to be present or quantifiable in retinal images, such as age (mean absolute error within 3.26 years), gender (area under the receiver operating characteristic curve = 0.97), smoking status (AUC = 0.71), systolic blood pressure (mean absolute error within 11.23 mmHg) and major adverse cardiac events (AUC = 0.70) (Fig. 1.5) [9].

The COVID-19 pandemic has raised expectations for the use of AI in data analysis. So far it has been used in epidemic modeling, detection of misinformation, diagnostics, vaccine and drug development, triage and patient outcomes, and identification of regions of greatest need [10].

Regulating AI-Based Medical Devices: Demonstrating Benefit and Safety

One of many challenges in the field of AI is determining what constitutes evidence of impact and benefit for AI medical devices and who should assess the evidence [2]. The majority of AI studies are conducted in experimental conditions and based on preselected data. They might provide inadequate insight into the use of AI applications in heterogeneous, real-world care settings.

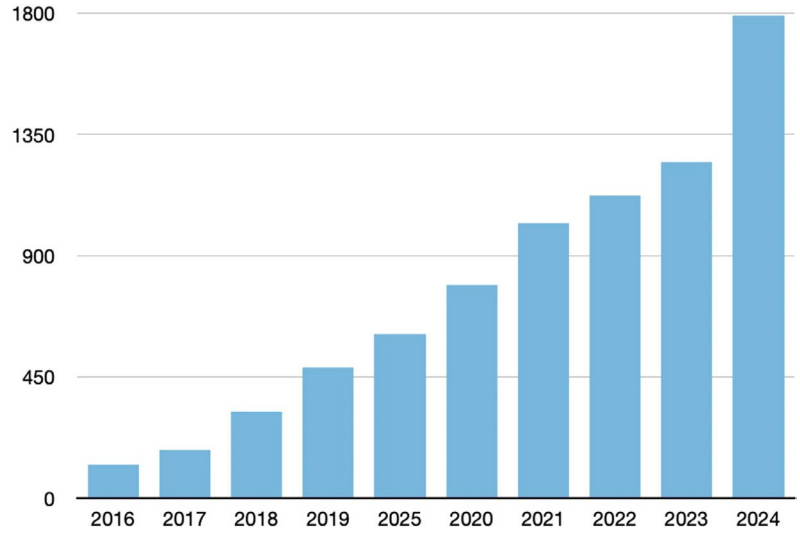

Fig. 1.3 The number of PubMed articles on Artificial Intelligence (AI) and the eye that were published between 2016 and 2024

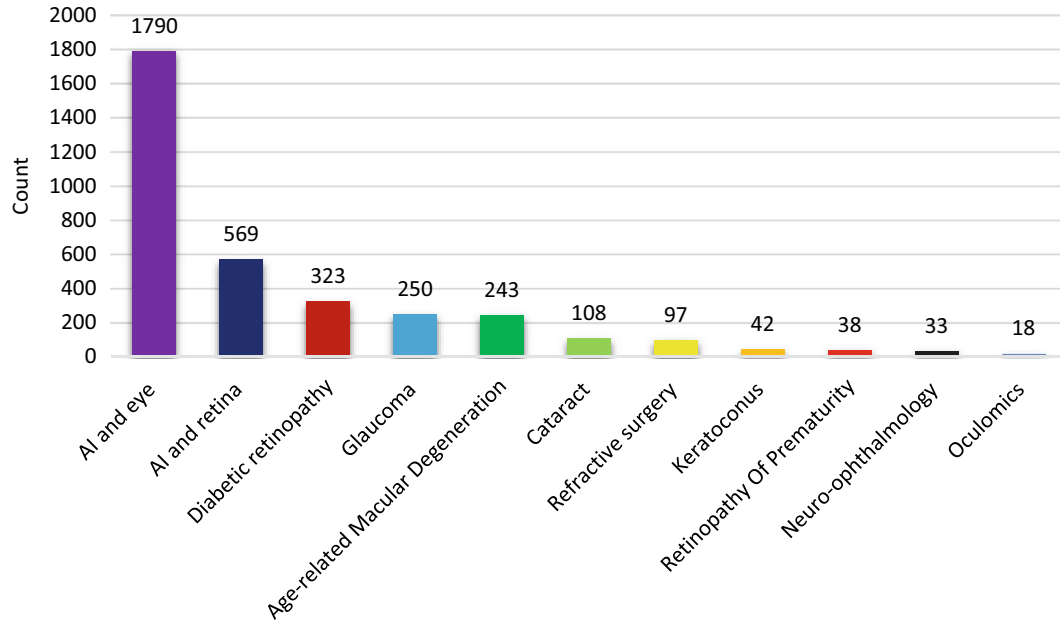

Fig. 1.4 The number of Pubmed articles relating to AI and eye diseases published in 2024

Lee et al. tested seven algorithms being used clinically around the world, including one with US Food and Drug Administration (FDA) approval and four whose developers have submitted applications for FDA approval. They found that most of these algorithms performed worse in real-world, compared with experimental, situations, with only three of seven and one of seven having comparable sensitivity and specificity to the human graders, respectively. Only one algorithm performed as well as human graders [11]. Another of the algorithms tested performed significantly worse than human graders at all levels of DR severity—it missed 25.58% of cases of advanced retinopathy, which could have serious consequences. One of the potential hazards of the clinical use of algorithms identified in this study was the risk of applying an algorithm trained on a particular demographic group to a population that differs in factors such as ethnicity, age, and sex. Moreover, many studies of algorithms developed with AI have excluded low-quality images, treating them as ungradable images, and patients with comorbid eye diseases, making them less reflective of real-world conditions.

The study by Lee et al. shows the importance and limitations of the registration process of AI-based medical devices. FDA registration is based on a centralized system, which does not have a specific, easily accessible regulatory pathway for AI-based medical devices. FDA clears the medical devices through three pathways: the premarket approval pathway, the de-novo premarket review, and the 510(k) pathway [12, 13]. The leading AI disciplines in medicine are radiology, cardiology, internal medicine/endocrinology, neurology, ophthalmology, emergency medicine, and oncology. FDA approvals of AI-based medical devices have increased steadily in recent years; there were 9 in 2015, 13 in 2016, 32 in 2017, 67 in 2018, and 77 in 2019, with the majority of devices designed for use in radiology, cardiology, and neurology [12]. Interestingly, 85% of FDA-approved medical devices in the years 2015–2019 were intended for use by health-care professionals, and only 15% for use by patients. The best-known, FDA-approved, AI-based medical devices in the field of ophthalmology are IDx-DR (2018), the first software to provide screening decisions that do not have to be interpreted by a clinician, and

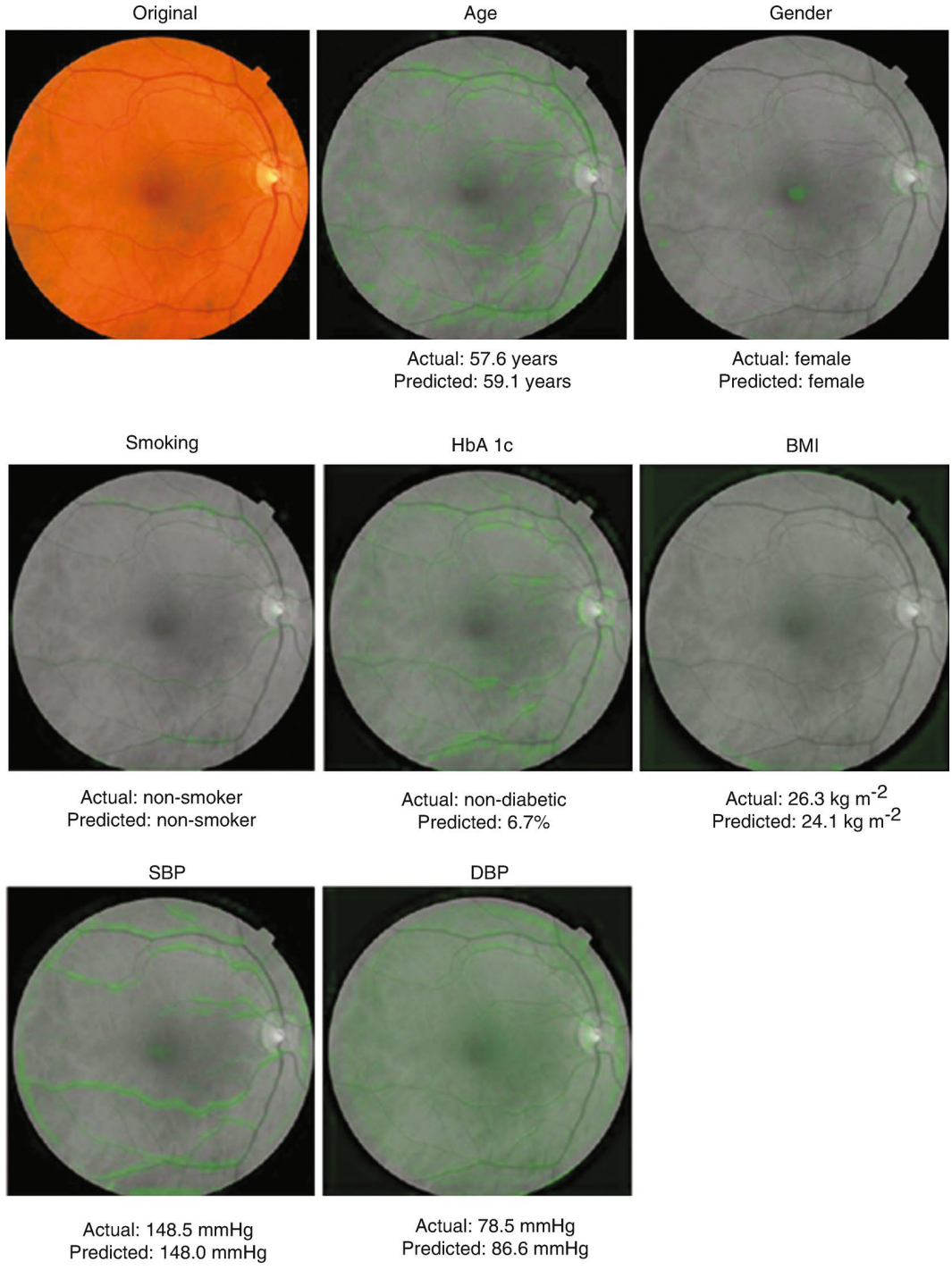

Fig. 1.5 Attention maps for a single retinal fundus image. The top left image is a sample retinal image in color from the UK Biobank data set. The remaining images show the same retinal image in black-and-white. The soft attention heat map for each prediction is overlaid in green, indicating the areas of the heat map that the neural-network model is using to make the prediction for the image. Source: Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018 Mar;2(3):158–164

Eyenuk (2020), which, like IDx-DR, screens for diabetic retinopathy.

In European Economic Area, which includes the European Union (EU) countries and the European Free Trade Association (EFTA) members (Iceland, Lichtenstein, Norway, and Switzerland), medical devices are approved in a decentralized manner.

Conformité Européenne (CE) marking indicates conformity with EU health, safety, and environmental-protection standards. For the lowest-risk medical devices (CE class I), the manufacturer ensures that the product complies with regulations and an approval procedure is not required. The registration procedure for higher-risk medical devices (CE class IIa, IIb, and III) is handled by private entities, called notified bodies, that have been accredited to assess the devices and issue a CE mark.

Thirteen CE-marked AI-based medical devices were approved in 2015, 27 in 2016, 26 in 2017, 55 in 2018, and 100 in 2019. The majority were designed for use in radiology, general hospital care, cardiology, neurology, ophthalmology (12 devices), and pathology, and most were class IIa (40%), class I (35%), or class IIb (12%) devices [12]. Of the AI-based devices that were CE-marked between 2015 and 2019, 124 (52%) were also FDA approved, making up 56% of the AI-based tools that the FDA approved in those. Bigger companies were more likely to get both approvals, whereas smaller companies were more likely to obtain only a CE mark. The authors of this study suggested that the European approval system was less rigorous than the US one. This conclusion is supported by an FDA report on 12 devices that received CE approval only and later were found to be unsafe or ineffective [13, 14]. A major problem in studying CE-marked devices in the European Economic Area is the lack of a publicly available register of approved devices comparable to the FDA register. Moreover, the information submitted to the notified bodies is confidential. In 2022, a new European database on medical devices (Eudamed), providing a live picture of the lifecycle of medical devices, will become operational. It will be composed of six modules, including actor registration, unique device identification (UDI), device registration, notified bodies and certificates, clinical investigations and performance studies, and vigilance and market surveillance [15].

Access to Reliable Data

DL algorithm training requires large data sets with thousands or even hundreds of thousands of diverse, well-balanced, and accurately labeled images [16]. The resources required for an AI study are presented in Fig. 1.6. The enormous numbers of required images rarely can be obtained from individual centers; thus they are secured from data repositories or centers that agree to share data. There is a growing need for consensus on standardized definitions of medical entities; conventions for data formatting; identification of units of measure; protocols for data cleaning, harmonization, and validation; standards for sharing and reusing data and sharing of code implementing AI models; and the adoption of open application program interfaces to AI models [17]. This is required for data sharing and open communication in AI, which is critical for conducting the reproducible research that is necessary before AI technology can be adopted in health care.

Kermany et al. used a DL analysis of a data set of optical coherence tomography images for triage and diagnosis of choroidal neovascularization, diabetic macular edema, and drusen. They demonstrated performance comparable to that of human experts and provided a more transparent and interpretable diagnosis by highlighting the regions recognized by the neural network. Further, they showed that a transfer-learning approach produced only modestly worse results (twofold increase of error, compared with the full data set) while using approximately 20 times fewer images. They also demonstrated the wider utility of this approach by applying it to the identification of pediatric pneumonia using chest X-ray images. They provided their data and code in a publicly available database to facilitate their use by other

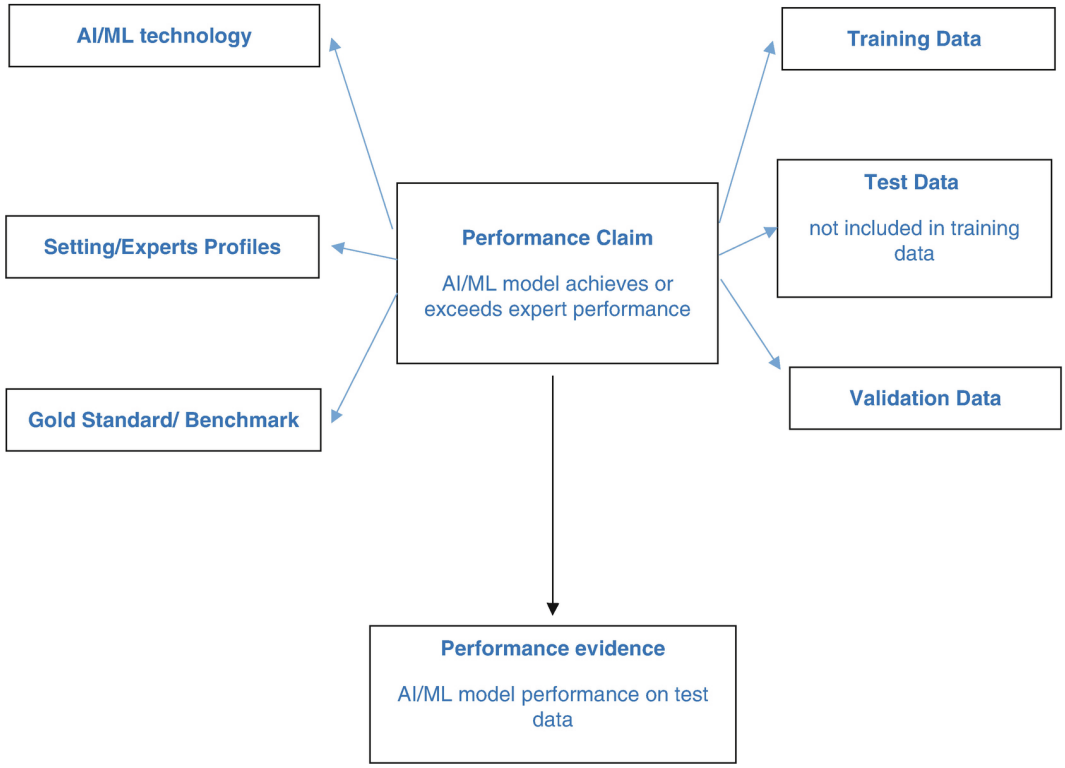

Fig. 1.6 The schematic presentation of resources required for an AI study

biomedical researchers in order to improve the performance of future models [18].

Transfer learning (Figs. 1.7 and 1.8) has been used in recent years to build classification models for medical images because the number of images that can be used for training is relatively small compared to the number of images available to train general models [19] (Fig. 1.9). Another approach to meet the need for large, annotated training data sets might be the use of low-shot DL algorithms. Low-shot learning (LSL), also known as few-shot learning, is a type of machine learning (ML) problems where the training dataset contains limited information. It is well known that many real-life situations, including rare diseases (e.g., serpiginous choroidopathy or angioid streaks in pseudoxanthoma elasticum) and non-typical presentations or subtypes of common disorders, are prone to AI bias due to the paucity or imbalance of data.

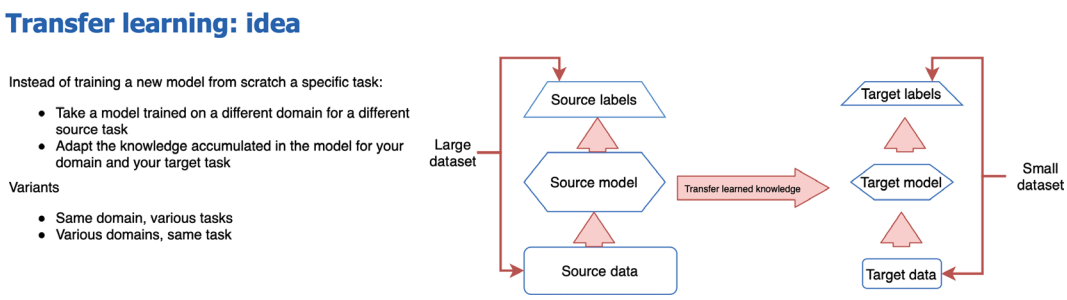

Fig. 1.7 The idea of transfer learning

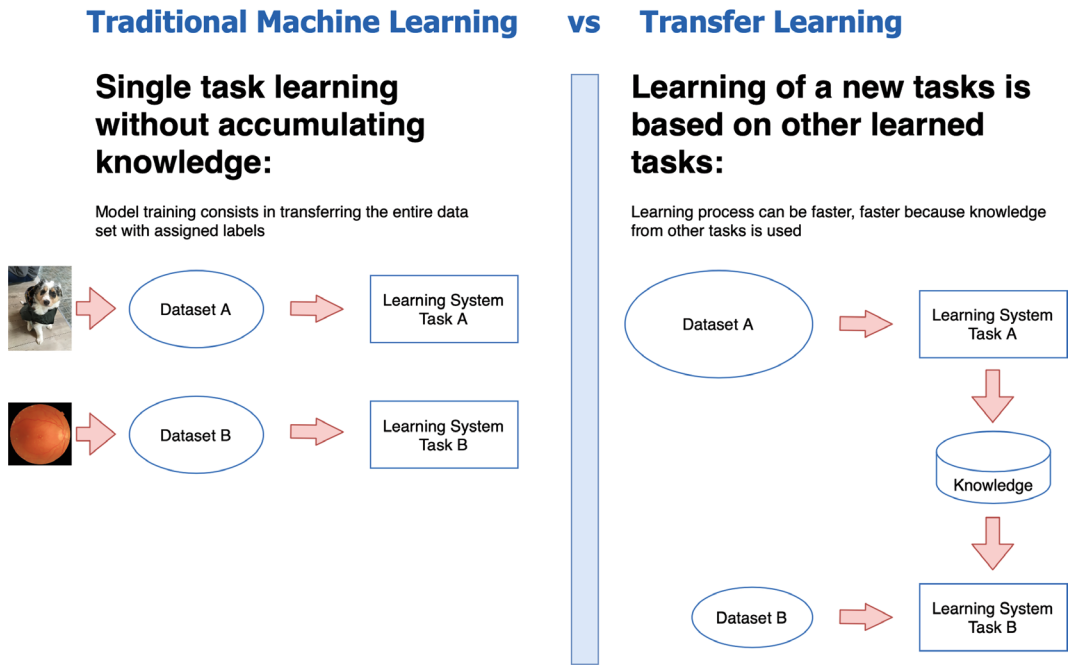

Fig. 1.8 Schematic of a convolutional neural network and transfer learning

These deficiencies may also result in less accurate future models. When addressing this sort of bias, dividing data according to some patient features (e.g., age, sex, and race/ethnicity) may result in smaller data sets that may be insufficient for training models for these particular groups. The study by Burlina et al. showed that the performance of widely used DL methods degraded substantially when used with limited data sets, but LSL methods performed better and might be applied in retinal diagnostics when a limited number of retina images are available for training [20].

Another approach that has been suggested by several authors to address the problem of limited data sets is the use of generative adversarial networks (GANs) to synthesize new images from a training data set of real images. GANs are ML models that can generate new data with the same statistics as the training set (Fig. 1.10). For example, a GAN trained on photographs can generate photographs of non-existing persons that look as authentic as real humans (Fig. 1.11). Artificial photos can be found at https://thispersondoesnotexist.com. Many applications of GAN have been proposed, including, art, fashion, advertising, science, video games, however, concerns about malicious uses were also raised, e.g., to produce fake, possibly incriminating, photographs and videos.

Burlina et al. used the Age-Related Eye Disease Study data set of over 130,000 fundus images to generate a similar number of synthetic images to train DL models. The performance of DL models trained with the synthetic images was nearly as good as the performance of models trained on real images [21]. Liu et al. have shown that 92% of synthetic OCT images had sufficient quality for further clinical interpretation. Only about 26–30% of synthetic post-therapeutic images could be accurately identified as synthetic images (Fig. 1.8) [22]. The accuracy of models trained on synthetic images to predict wet or dry macular status was 0.85 (95% CI 0.74–0.95) [22]. In a study by Zheng et al., the image quality of real versus synthetic OCT images was similar as assessed by two retinal specialists. The accuracy of discrimination of real versus synthetic OCT images was 59.50% for retinal specialist 1 and 53.67% for retinal

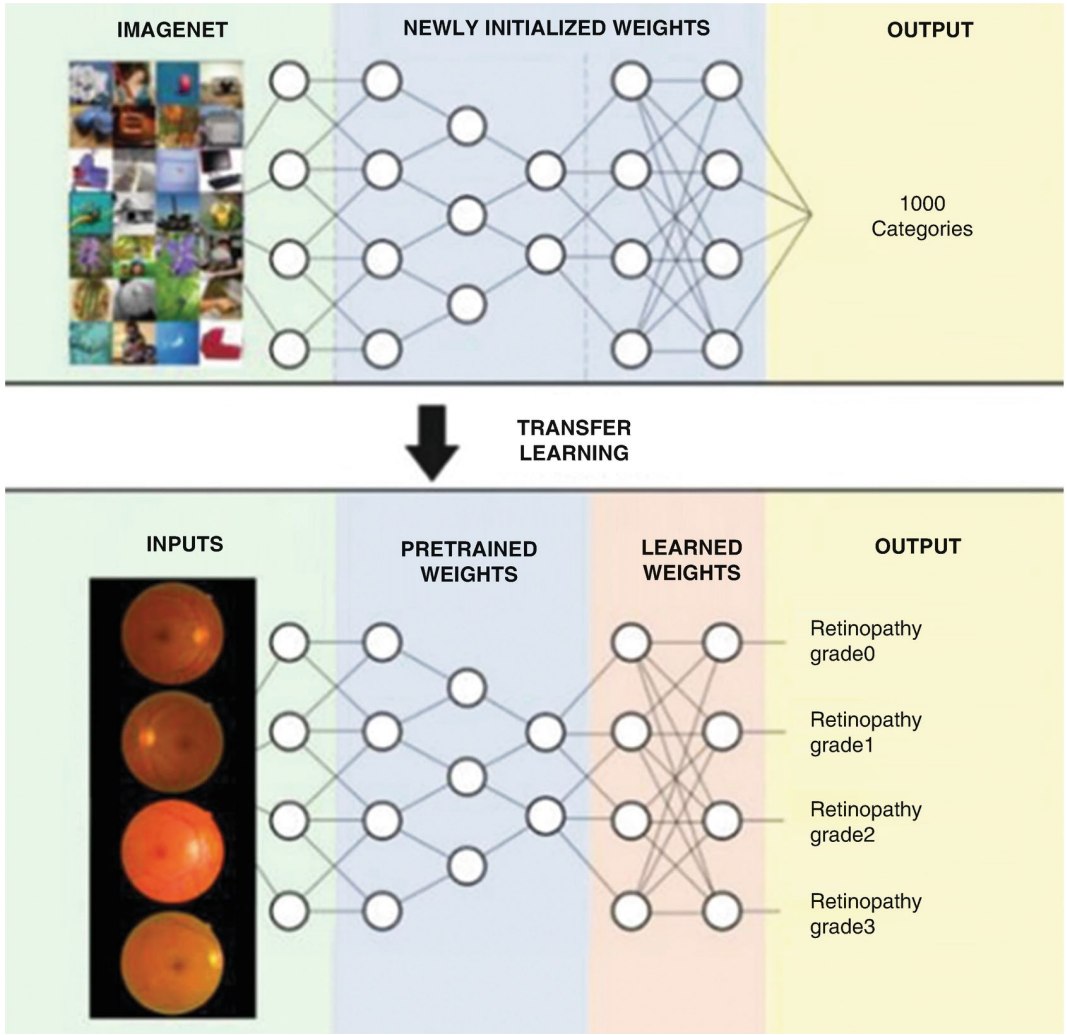

Fig. 1.9 The schematic diagram of transfer learning. Source: Lingling Li et al. Diabetic retinopathy identification system based on transfer learning. 2020. *J. Phys.: Conf. Ser.* 1544 012,133. https://doi.org/. Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence

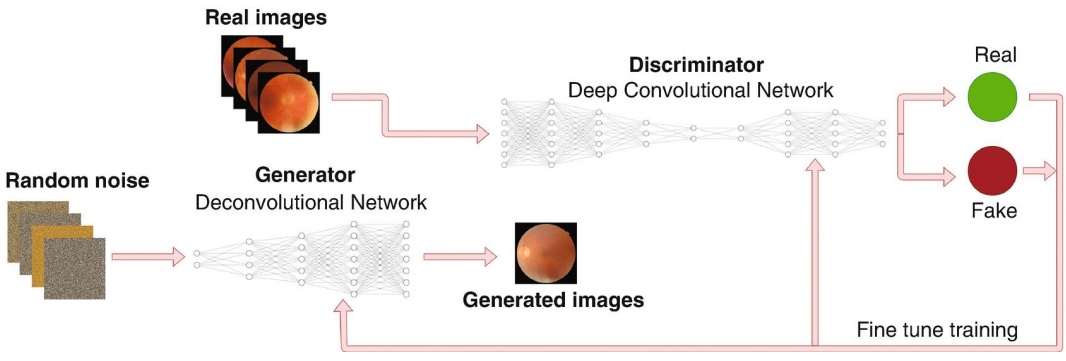

Fig. 1.10 The schematic presentation of generative adversarial network (GAN)

Fig. 1.11 The image of a young woman generated by StyleGAN, an generative adversarial network (GAN). The person in this photo does not exist, but is generated by an artificial intelligence based on an analysis of portraits. Source: https://commons.wikimedia.org/wiki/File:Woman_1.jpg. This file is in the public domain because, as the work of a computer algorithm or artificial intelligence, it has no human author in whom copyright is vested

specialist 2. For the local data set, the DL model trained on real and synthetic OCT images had an area under the curve of 0.99 and 0.98, respectively. For the clinical data set, the area under the curve was 0.94 for the real model and 0.90 for the synthetic one [23]. These studies suggest the GAN synthetic images can be used by clinicians for educational purposes and developing DL algorithms [24].

An important and interesting issue is the clinical application of continual ML, i.e., continuous learning and development from new data while retaining previously learned knowledge [25]. However, there are technical challenges to the implementation of this promising concept, including the need to prevent interference between new and old data, and old and new knowledge. In the catastrophic interference phenomenon, the acquisition of new data can lead to an abrupt decrease in the performance of an algorithm. Practical applications of AI tools in health care must be cautiously introduced because of the existence of such risks. FDA regulations require that FDA-approved autonomic algorithms be locked for safety to prevent unpredictable future changes. This requirement, however, is designed to ensure the safety of the model rather than improving its performance. Continual learning could refine the performance of machine-learning algorithms by the gradual correction and elimination of mistakes. It will be necessary to consider how this technology can be introduced safely to health care.

Hazards and Challenges of AI in Ophthalmology

The future development of ophthalmology depends on better and possibly unlimited access to the medical data stored within electronic health records. However, this access cannot be allowed to compromise of privacy of this very sensitive data. There is a need for effective regulations that will set a balance between individual protection and the common good. One approach to protecting privacy and increasing sample size is to share DL algorithms with local institutions for retraining purposes, but without sharing the private data used to build the algorithms. This model-to-data approach, also known as federated learning, was tested in ophthalmology and was shown to work effectively [26].

According to the US National Institute of Standards and Technology, biometric data, including retina images, are personally identifiable information and should be protected from inappropriate access. Although AI models have been shown to diagnose and stage some ocular diseases from fundus photographs, OCT, and visual-field images, most AI algorithms were tested on data sets that did not correspond well to real-world conditions. Patient populations were usually homogenous, and poor-quality images and patients with multiple pathologies were excluded. Future studies are needed to validate algorithms on ocular images from heterogeneous populations, including both good- and poor-quality images. Otherwise, we may face the situation of "good AI gone bad." The tendency to cherry-pick the best results might

make the situation even worse. AI algorithms can behave unpredictably when applied in real life. Algorithm performance can degrade after deployment due to the changes between the training and testing conditions (dataset shift), caused, for example, by using to images generated by a different device than this in the training set or collected in a different clinical environment [27–30]. Moreover, algorithms may return different outputs at different times when presented with similar inputs [31, 32]—they can be affected by minor changes in image quality or extraneous data on an image [32–35]. All these problems might lead to misdiagnosis and erroneous treatment suggestions, breaching trust in AI technologies. An error in an AI system could harm hundreds or even thousands of patients.

A recent report from the National Academy of Medicine [36] highlights some important challenges in the further development of AI applications in health care (Table 1.2). Its authors advocate the use of openly accessible, standardized, population-representative data; addressing explicit and implicit biases related to AI; developing and deploying appropriate training and educational programs for health workers to support health-care AI; and balancing innovation and safety through the use of regulation and legislation to promote trust.

To understand the limitations of AI-based models in health care and the responsibilities of manufacturers and users of AI software as a medical device (SaMD), an MI-CLAIM checklist was proposed for use in AI software development [37]. Its purpose is to enable a direct assessment of clinical impact, including considerations of fairness and bias, and to allow rapid replication of the technical design by any legitimate clinical AI study. The MI-CLAIM checklist has six parts (Table 1.3), including (1) Study design; (2) Separation of data into partitions for model training and model testing; (3) Optimization and final model selection;

Table 1.2 Practical challenges to the advancement and application of AI tools in clinical settings

Workflow integration	Understand the technical, cognitive, social, and political factors in play and incentives impacting integration of AI into health care workflows.
Enhanced explainability and interpretability	To promote integration of AI into health care workflows, consider what needs to be explained and approaches for ensuring understanding by all members of the health care team.
Workforce education	Promote educational programs to inform clinicians about AI/machine learning approaches and to develop an adequate workforce.
Oversight and regulation	Consider the appropriate regulatory mechanism for AI/machine learning and approaches for evaluating algorithms and their impact.
Problem identification and prioritization	Catalog the different areas of health care and public health where AI/machine learning could make a difference, focusing on intervention-driven AI
Clinician and patient engagement	Understand the appropriate approaches for involving consumers and clinicians in AI/machine learning prioritization, development, and integration, and the potential impact of AI/machine learning algorithms on the patient-provider relationship.
Data quality and access	Promoting data quality, access, and sharing, as well as the use of both structured and unstructured data and the integration of non-clinical data is critical to developing effective AI tools.

Source Matheny ME, Thadaney Israni S, Ahmed M, Whicher D. AI in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC: National Academy of Medicine; 2019. https://nam.edu/artificial-intelligence-special-publication

Table 1.3 The MI-CLAIM checklist [Source: Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020 Sep;26(9):1320–1324]

Before paper submission		
Study design (Part 1)	*Completed: page number*	*Notes if not completed*
The clinical problem in which the model will be employed is clearly detailed in the paper	□	
The research question is clearly stated	□	
The characteristics of the cohorts (training and test sets) are detailed in the text	□	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings	□	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed	□	
Data and optimization (Parts 2, 3)	*Completed: page number*	*Notes if not completed*
The origin of the data is described and the original format is detailed in the paper	□	
Transformations of the data before it is applied to the proposed model are described	□	
The independence between training and test sets has been proven in the paper	□	
Details on the models that were evaluated and the code developed to select the best model are provided	□	
Is the input data type structured or unstructured?	□	
Model performance (Part 4)	*Completed: page number*	*Notes if not completed*
The primary metric selected to evaluate algorithm performance (e.g., AUC, F-score, etc.), including the justification for selection, has been clearly stated	□	
The primary metric selected to evaluate the clinical utility of the model (e.g., PPV, NNT, etc.), including the justification for selection, has been clearly stated	□	
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance	□	
Model examination (Part 5)	*Completed: page number*	*Notes if not completed*
Examination technique 1[a]	□	
Examination technique 2[a]	□	
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented	□	
A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented	□	
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included	□	
Reproducibility (Part 6): choose appropriate tier of transparency		*Notes*
Tier 1: complete sharing of the code	□	
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation	□	
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details	□	

(continued)

Table 1.3 (continued)

Before paper submission		
Tier 4: no sharing	□	

PPV positive predictive value, *NNT* numbers needed to treat
[a]Common examination approaches based on study type: for studies involving exclusively structured data, coefficients and sensitivity analysis are often appropriate; for studies involving unstructured data in the domains of image analysis or natural language processing, saliency maps (or equivalents) and sensitivity analyses are often appropriate

Table 1.4 Major topics of CONSORT-AI extension

1. State the inclusion and exclusion criteria at the level of participants
2. State the inclusion and exclusion criteria at the level of the input data
3. Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements
4. State which version of the AI algorithm was used
5. Describe how the input data were acquired and selected for the AI intervention
6. Describe how poor-quality or unavailable input data were assessed and handled
7. Specify whether there was human—AI interaction in the handling of the input data, and what level of expertise was required for users
8. Specify the output of the AI intervention
9. Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice
10. Describe results of any analysis of performance errors and how errors were identified, where available. If no such analysis was planned or done, explain why not.

Source Adapted from Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020 Oct;2(10):e537–e548

(4) Performance evaluation; (5) Model examination; and (6) Reproducible pipeline. The CONSORT-AI and SPIRIT-AI working groups have proposed reporting guidelines for clinical trials of interventions involving AI. A summary of these guidelines is presented in Table 1.4.

Inherent conflicts of interest should be acknowledged. Manufacturers who develop and market SaMD have a strong financial interest in presenting their products positively. Thus, conflicts of interest exist if they fund, conduct, and publish results of studies, including those that might report deficiencies in their products. Many of the published papers in the field of AI-based diabetic retinopathy screening, particularly those using CE-marked and FDA-approved algorithms, were conducted by manufacturers or patent owners.

It should be also remembered that AI algorithms can be designed to perform in unethical ways. For example, Uber's software Greyball allowed the company to identify and circumvent local regulations, and Volkswagen's algorithm allowed vehicles to pass emission tests by reducing their emissions of nitrogen oxide during testing. AI algorithms could be tuned to generate increased profits for their owners by recommending particular drugs, tests, or the like without clinical users' awareness. AI systems are vulnerable to cybersecurity attacks that could cause their algorithms to misclassify medical information [31].

Seven essential factors to design AI for social good were proposed by Floridi et al. (Table 1.5) [38]. The authors propose falsifiability as an essential factor to improve the trustworthiness of technological application, i.e., for an SaMD to be trustworthy, its safety should be falsifiable. Critical requirements for a device to be fully functional must be specified and must be testable. If falsifiability is not possible, then the critical requirements cannot be checked, and the system should not be deemed trustworthy [38].

Cost-Effectiveness of AI-Based Devices

One of the arguments for AI-based medical devices is that they can reduce medical costs and eliminate unnecessary procedures. A study from Singapore found that a semiautomated model that

Table 1.5 Essential Factors to design AI for social good

Factors	Corresponding best practices	Corresponding ethical principle
Falsifiability and incremental deployment	Identify falsifiable requirements and test them in incremental steps from the lab to the "outside world"	Nonmaleficence
Safeguards against the manipulation of predictors	Adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation	Nonmaleficence
Receiver-contextualised intervention	Build decision-making systems in consultation with users interacting with and impacted by these systems; with understanding of users' characteristics, the methods of coordination, the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions	Autonomy
Receiver-contextualised explanation and transparent purposes	Choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers; then deploy arguments that are rationally and suitably persuasive for the receiver to deliver the explanation; and ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default	Explicability
Privacy protection and data subject consent	Respect the threshold of consent established for the processing of datasets of personal data	Nonmaleficence; autonomy
Situational fairness	Remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives	Justice
Human-friendly semanticisation	Do not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something	Autonomy

Source Floridi L, Cowls J, King TC, Taddeo M. How to Design AI for Social Good: Seven Essential Factors. Sci Eng Ethics. 2020 Jun;26(3):1771–1796. Springer

combined a DL system with human assessment achieved the best economic returns, leading to savings of 19.5% in screening for diabetic retinopathy. An earlier study from the UK reported cost-savings of 12.8–21.0%; however, a simple comparison between them is not possible due to the different models of DR screening in the two countries (two-stage screening in Singapore and three-stage screening in the UK), and their different DR classification systems. The authors of both studies argued that a semiautomated system produces more savings than a fully automated system due to the lower rate of false positives and unnecessary specialist visits [39, 40].

This book aims to provide ophthalmologists and other visual professionals and researchers with an overview of current research into the use of AI in ophthalmology. Together with a team

of international experts from Europe, North America, and Asia, we present an overview of the most important documentary research in ophthalmology on ML and AI technologies and their benefits. We discussed the use of AI in the diagnosis of some retinal and corneal disorders, the diagnosis of congenital cataract, neuro-ophthalmology, glaucoma, intraocular lens calculation methods, ocular oncology, ophthalmology triaging, cataract-surgery training, refractive surgery, and the assessment and prediction of systemic diseases through the use of the eye. Chapters on digital-image analysis, AI basics, and technical aspects of AI provide the reader with knowledge not commonly possessed by ophthalmologists, but required to understand the topic in both its field-specific and broader contexts. The very important chapter on AI safety and efficacy outlines the challenges ophthalmology will face with the introduction and widespread dissemination of this technology. Although we have covered all of the major areas of AI/ML technology in ophthalmology, research in this field is progressing so quickly that some new concepts that emerged at the end of 2020 and in early 2021 do not appear on these pages. However, evidence-based medicine often demands that we await for more evidence to verify early reports and assess the real value of new medical technologies or applications. I would like to thank all the contributors for sharing their knowledge in this new and fascinating discipline, which has great potential to change ophthalmology.

Acknowledgements I would like to thank Aleksandra Lemanik, Foundation for Ophthalmology Development, Poznan, Poland and Tomasz Krzywicki, Faculty of Mathematics and Computer Science, University of Warmia and Mazury, Olsztyn, Poland for their help in preparing illustrations, and Szymon Wilk, Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan, Poland, for his valuable discussion on this chapter.

References

1. Mitchell M. Artificial intelligence: a guide for thinking humans. Penguin UK; 2019.
2. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books; 2019.
3. Copeland BJ. Artificial intelligence. *Encyclopedia Britannica*, 11 August 2020. https://www.britannica.com/technology/artificial-intelligence. Accessed 18 Mar 2021.
4. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf. 2014;23(9):727–31.
5. Gunderson CG, Bilan VP, Holleck JL, et al. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. BMJ Qual Saf. 2020;29:1008–18.
6. Zwaan L, Singh H. Diagnostic error in hospitals: finding forests not just the big trees. BMJ Qual Saf. 2020;29(12):961–4.
7. Singletary B, Patel N, Heslin M. Patient perceptions about their physician in 2 words: the good, the bad, and the ugly. JAMA Surg. 2017;152(12):1169–70.
8. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3:118.
9. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158–64.
10. Chen J, See KC. Artificial intelligence for COVID-19: rapid review. J Med Internet Res. 2020;22(10):e21476.
11. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, Gencarella MD, Gee H, Maa AY, Cockerham GC, Lynch M, Boyko EJ. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care. 2021;dc201877. https://doi.org/10.2337/dc20-1877.
12. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. 2021;3(3):e195-203.
13. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. JAMA. 2019;322(23):2285–6.

14. Hwang TJ, Sokolov E, Franklin JM, Kesselheim AS. Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. BMJ. 2016;353: i3323.
15. European Commission. Medical devices—EUDAMED. 17 June 2020. https://ec.europa.eu/growth/sectors/medical-devices/new-regulations/eudamed_en. Accessed 15 Jan 2021.
16. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med. 2018;24(5):539–40.
17. Wang SY, Pershing S, Lee AY, AAO Taskforce on AI and AAO Medical Information Technology Committee. Big data requirements for artificial intelligence. Curr Opin Ophthalmol. 2020;31(5):318–23.
18. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122-1131.e9.
19. Rampasek L, Goldenberg A. Learning from everyday images enables expert-like diagnosis of retinal diseases. Cell. 2018;172(5):893–5.
20. Burlina P, Paul W, Mathew P, Joshi N, Pacheco KD, Bressler NM. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. JAMA Ophthalmol. 2020;138(10):1070–7.
21. Burlina PM, Joshi N, Pacheco KD, Liu TYA, Bressler NM. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. JAMA Ophthalmol. 2019;137:258–64.
22. Liu Y, Yang J, Zhou Y, Wang W, Zhao J, Yu W, Zhang D, Ding D, Li X, Chen Y. Prediction of OCT images of short-term response to anti-VEGF treatment for neovascular age-related macular degeneration using generative adversarial network. Br J Ophthalmol. 2020;104(12):1735–40.
23. Zheng C, Xie X, Zhou K, Chen B, Chen J, Ye H, Li W, Qiao T, Gao S, Yang J, Liu J. Assessment of generative adversarial networks model for synthetic optical coherence tomography images of retinal disorders. Transl Vis Sci Technol. 2020;9(2):29.
24. Liu TYA, Farsiu S, Ting DS. Generative adversarial networks to predict treatment response for neovascular age-related macular degeneration: interesting, but is it useful? Br J Ophthalmol. 2020;104(12):1629–30.
25. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. 2020;2(6):e279–81.
26. Mehta N, Lee CS, Mendonça LSM, Raza K, Braun PX, Duker JS, Waheed NK, Lee AY. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. JAMA Ophthalmol. 2020;138(10):1017–24.
27. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. J Am Coll Radiol. 2021;18(3 Pt A):413–24.
28. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. J Am Coll Radiol. 2020;17:796–803.
29. Subbaswamy A, Schulam P, Saria S. Preventing failures due to dataset shift: learning predictive models that transport. Proc Mach Learn Res. 2019;89:3118–27.
30. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21:345–52.
31. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195.
32. Winkler JK, Fink C, Toberer F. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. 2019;155:1135–41.
33. Finlayson SG, Bowers JD, Ito J. Adversarial attacks on medical machine learning. Science. 2019;363:1287–9.
34. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15: e1002683.
35. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci U S A. 2020; pii:201907377. https://doi.org/10.1073/pnas.1907377117.
36. Matheny ME, Thadaney Israni S, Ahmed M, Whicher D. AI in health care: the hope, the hype, the promise, the peril. Washington, DC: National Academy of Medicine; 2019. https://nam.edu/artificial-intelligence-special-publication
37. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320–4.
38. Floridi L, Cowls J, King TC, Taddeo M. How to design AI for social good: seven essential factors. Sci Eng Ethics. 2020;26(3):1771–96.

39. Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, Yip MYT, Qi Lee X, Hsu W, Li Lee M, Tan CS, Tym Wong H, Lamoureux EL, Tan GSW, Wong TY, Finkelstein EA, Ting DSW. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health. 2020;2(5):e240–9.

40. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, Lee A, Louw V, Anderson J, Liew G, Bolter L, Srinivas S, Nittala M, Sadda S, Taylor P, Rudnicka AR. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology. 2017;124(3):343–51.

Basics of Artificial Intelligence for Ophthalmologists

2

Ikram Issarti and Jos J. Rozema

Abstract

The applications of Artificial Intelligence are slowly beginning to take over tasks into the field of ophthalmology that until recently required a trained clinician to perform. Given their exceedingly level of high performance, many are inclined to trust the results provided by such programs, without much insight into how these results were accomplished. The exponential advancement of AI, especially in topics such as Natural Language Processing (NLP) and Generative AI expanded the use cases and the capabilities of AI, triggering boarder concerns and capabilies of the impact and usability of such technologies in the area of Ophtalmology. This chapter comprehensively elucidates the basic principles and latest updates of Artificial Intelligence to help ophthalmologists get a feel for the strengths and limits of the techniques, as well as how they may use them in clinical practice.

Keywords

Artificial intelligence · Machine learning · Deep learning · Ophthalmology · Introduction · Basic principles · Generative AI · Large language models

Introduction

The past decade has seen a steep rise in the number of applications of Artificial Intelligence (AI), especially for repetitive or complex tasks where humans may quickly suffer from either a drifting attention span or subtle inconsistencies. Such systems are often more cost efficient, thus accelerating their adoption and acceptance, consequently increasing people's reliance on AI. But an understanding of its inner workings is often lacking, many tend to approach it as a 'black box' at the risk of uncritically accepting whatever output it produces. Although by its very nature, AI is opaque about how it reaches a result, there are statistical methods to objectively assess the quality of its output. As AI becomes a popular subject within the scientific community and health-care practice, this chapter explains the basic principles of AI in a comprehensive step-by-step manner, along with examples of ophthalmological applications. Special

I. Issarti · J. J. Rozema (✉)
Visual Optics Lab Antwerp (VOLANTIS), Faculty of Medicine and Health Sciences, University of Antwerp, Wilrijk, Belgium
e-mail: jos.rozema@uantwerpen.be

Department of Ophthalmology, Antwerp University Hospital, Edegem, Belgium

A. Grzybowski (ed.), *Artificial Intelligence in Ophthalmology*,
https://doi.org/10.1007/978-3-031-83756-2_2

attention will be paid to the differences between AI, Machine Learning (ML), and Deep Learning (DL), highly interconnected techniques that are often confused for one another. This work's scope will extend to cover recent trends in AI including Generative AI and Large Language Models, together with certain recent applications in Ophtalmology.

Short History

The concept of artificial intelligence dates back to ancient times, with legends describing artificial beings created by master craftsmen or brought into life by magic. For many centuries, and without the required scientific and technological basis, the aspiration of creating intelligent beings remained fictitious and only present in art, myths, and novels.

It was not until the eighteenth century that the mathematical and statistical foundations of AI were established, such as Bayes' Theorem, which would later become the groundwork for the current AI developments. Simultaneously, the first executable algorithms were developed by Ada Lovelace, who argued that computers could be used to execute instructions beyond basic calculations.

In the nineteenth century, Neuroscience has inspired the first establishment of the theoretical foundations of AI. Warren McCulloch and Walter Pitts introduced a computational Artificial Neuron [1], with the intent of explaining how the biological neuron wires. The computational Neuron functioning was described by simple logical operators (AND, OR, and NOT). A few years later, Donald Hebb published his famous paper "The Organization of Behavior" [2], in which he suggested that the biological neuronal pathways are reinforced with repeated use. Hebb's work inspired the earliest AI algorithms, Artificial Neural Networks, which are assumed to mimic certain intelligente characteristics, such as learning, adaptability, and generalization.

From the 1950s onward, specific AI concepts have developed, yet the term Artificial Intelligence isn't yet known to the scientific public. At the time, philosophers, and scientists were exploring the idea of achieving machines that could 'Think' and perform tasks restricted to human intelligence. This was the scope of the summer workshop organized by John MacCarthy at Dartmouth College [3]. In the same vein, the famous English mathematician Alan Turing published his paper '*Computing Machinery and Intelligence*' [4], in which he argued that the focus may need to be shifted toward developing machines that could perform human-like tasks instead of focusing on achieving '*Thinking*' machines. Turing also developed his famous Turing test, which is an open-ended interview with a human evaluator that must determine whether the respondent is human or machine. Although AI has yet to pass the test even today [5], it helped create the concept of 'Artificial Intelligence' as a result of MacCarthy's conference.

Between 1956 and 1974, advancements in computing power enabled the testing of various AI concepts. Among them, the 'Logic Theorist' [6], represented the initial AI program that emulated human problem-solving abilities. Followed by Samuel's self-learning checkers-playing program. Thus, introducing the term 'Machine Learning' [7]. These critical milestones, however, marked a benchmark in the evolution of AI and machines, as learning algorithms emerged. Subsequently, a successful first development of the Natural Language Processing program ELIZA [8].

In spite of the initial successes, the 1970s witnessed the first AI Winter, due to limited funding, interest, and unmet expectations. It was concluded that if strong AI was not achievable, then other sciences could solve typical AI problems. Yet, notable AI breakthroughs continued. The first Optical Character Recognition (OCR) and text-to-speech synthesizer were developed, allowing computers to read text out loud for blind patients [9]. The following years saw a resurgence in AI activities, particularly 'Expert Systems', which were designed to capture human knowledge in specific domains and mimic the human decision-making process. Meanwhile, Artificial Neural Networks

(a subfield of AI), and backpropagation algorithms, witnessed significant progress which is considered the fundamental basis of Machine and Deep Learning today. This progress was joined with the development of parallel distributed processing, robotics, AI-medical diagnosis applications, etc. This continued up to the 90s, when AI progress recessed again in the second AI winter due to limitations in the computing power required for implementing robust expert systems. Even so, during this time the first 'Intelligent Agents' emerged, a broad term for an AI that can perceive its environment and respond accordingly, such as IBM's Deep Blue that defeated chess world champion Garry Kasparov or LeNet-5, a Convolutional Neural Network for digit recognition. Industrialized and marketed AI progress appeared from the early 2000s onwards due to a significant increase in the availability of datasets and computing power, leading to large breakthroughs in Machine Learning algorithms and applications. These include for instance the iRobot Roomba autonomous vacuum cleaner released in 2002 and the self-driving vehicles introduced by Google, Tesla and Ford.

In 2006, Deep Belief Networks were introduced, sparking more interest in the potential of Deep Learning. Various successful Deep Learning applications were developed, such as AlexNet, a Convolutional Neural Network that successfully competed in the ImageNet competition for image recognition. Similar breakthroughs included Google's Word2Vec, a Shallow Neural Network that represents words, as well as DeepMind's AlphaGo, which defeated the Go World Champion after only 30 hours of unsupervised learning.

In the following years the Machine Learning and Deep Learning algorithms expanded in scope and capabilities as dozens of new algorithms were developed and commercialized while the scientific community explored other types of AI, such as the Generative Adversarial Networks (GANs, 2014) [10] and dozens of subsequent variants. In the famous 2017 paper '*Attention is All you Need*' by Vaswani et al. [11], GANs were combined with a Transformer architecture, the key components behind Large Language Models like ChatGPT and BERT, or image generators like DALL-E and CLIP. Such algorithms have revolutionized the AI landscape by providing proven versatility without the need for fine-tuning. The development of Large Language Models is still ongoing, however, as OpenAI recently released GPT-4o, that will bridge into GPT-5 later in 2024.

Overview

The following gives an overview of key terminologies and algorithms employed in AI.

Artificial Intelligence is a very broad field of study encompassing a wide range of techniques that allow machines to display ever more intelligent behaviour (Fig. 2.1). **Machine learning** is one of the most important subfields of AI. Although ML and AI are often confused, AI also includes other approaches not included in machine learning, such as Expert Systems, knowledge - or rule-based systems that emulate human cognitive and reasoning abilities by following certain guidelines to perform a decision-making process [12]. Meanwhile, ML refers to a group of mathematical algorithms that learn from experience (data) by mimicking human learning behaviour to perform new tasks. ML is able to fit complex data sets, to extract new knowledge, imitate complex behaviour, predict and classify based on prior data. Another well-known group of algorithms is **Deep learning** (DL), which is a subset of machine learning based on artificial neural networks. DL is able to simultaneously analyse multiple layers of data. These layers consist of data processing units, called neurons, that allow them to analyse large amounts of data at once while preserving the data's spatial distribution. DL systems have seen significant successes in applications such as pattern recognition, image processing, and speech recognition. The application of DL in linguistics and semantics is referred to as **Natural Language Processing (NLP)**, a subset of DL that focuses on the interaction between computers and humans through natural language by interpreting

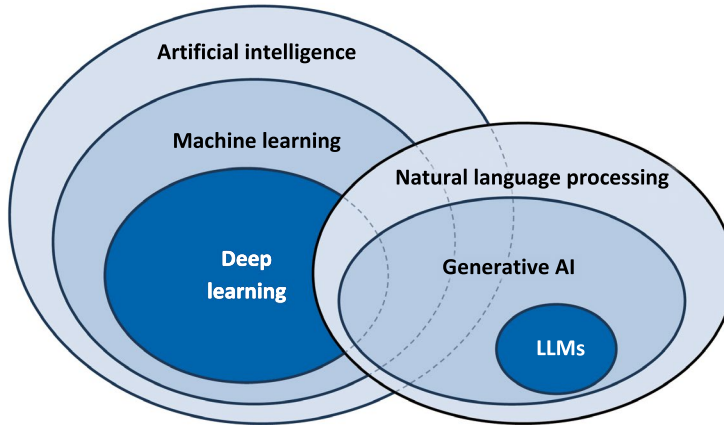

Fig. 2.1 Artificial intelligence techniques

and manipulating sequential semantics. It has been around since the 1960s, but due to recent advances in ML and DL, NLP has become the foundation for the disruptive digital applications being developed, ranging from automating clinical documentation to enhancing patient experience with chatbots. Building on the capabilities of NLP, **Generative AI** is able to learn to model the underlying distribution of a given dataset and subsequently generate new content, such as text, images, and music. Applications, may include the generation of new survey questions or synthetic medical data for research.

While the above-mentioned learning types (ML, DL, NLPs, and Generative AI) focus on specific tasks, they are still categorized as narrow AIs as they are solely for single-purpose applications. Strong AI or **Artificial General Intelligence (AGI)**, on the other hand, aspires to achieve human levels of intelligence, with intelligence features such as understanding, reasoning, and learning across various domains without re-training. Although strong AI is currently still a theoretical concept, the ongoing effort in developing large multimodal (combined) AI shows signs of approaching AGI as for instance **Large Language Models** (LLMs) have proven themselves to be a powerful general purpose platform capable of performing a wide range of tasks without additional training. These LLMs (e.g., GPT-4, BERT, etc.) are NLPs with an underlying Transformer architecture, trained on trillions of data by parallel supercomputers to interpret human language, generate text, and engage in human-like conversations. Unlike standard, data-driven AI (e.g., ML, DL), LLMs are trained on vast amounts of data and billions of parameters, which enables them to be general-purpose and perform a variety of tasks without fine-tuning. One example is ChatGPT, a Generative AI chatbot that generates human-like responses in a conversation through prompts and without the need for re-training or expert coding. But an LLM can also be fine-tuned on specific data to make it more relevant and avoid incorrect or inappropriate responses.

The **training process** of most AI algorithms (e.g., ML, DL, etc.) is very similar to that found in schools, with a professor teaching his students. From a large amount of given data, the algorithm learns how to describe a specific topic into a model (knowledge acquisition), which will subsequently be validated using unseen data to evaluate its generalizability. Finally, the performance of the algorithm is evaluated based on several guidelines given in the "Performance Evaluation" section.

Data Basis

Data is the fuel of AI, which can come from different sources such as, webs, videos, audios, text, etc. It is comprised of massive amounts of

bits, binary values of zeros and ones, that can be reorganized to form structured data that is usually easier to process by AI algorithms, such as a relational database or a spreadsheet. It is also possible to work with unstructured data without predefined formatting (e.g. audio, video, text, etc.), or with a hybrid form of a structured and unstructured data called semi-structured data. Finally, one can consider time series data, consisting of structured or unstructured data in sequential time steps [13]. A good understanding of data structures allows a proper AI implementation. Some highlights are given in the section "Conducting a Machine Learning Analysis", but more details are available in the data mining literature and data pre-processing text books [14].

Common Tasks

In medicine, Machine Learning is mostly used to assist physicians with diagnosis, monitoring, and decision making by providing insight into the structure and patterns within large datasets. The most typical tasks for ML are classification, clustering and prediction.

- **Classification** involves sorting new cases into two or more groups (Fig. 2.2a). In healthcare, classification could be used for diagnosis (healthy or abnormal) or the identification of biological markers.
- **Clustering** the algorithm divides a dataset into several, previously unknown clusters

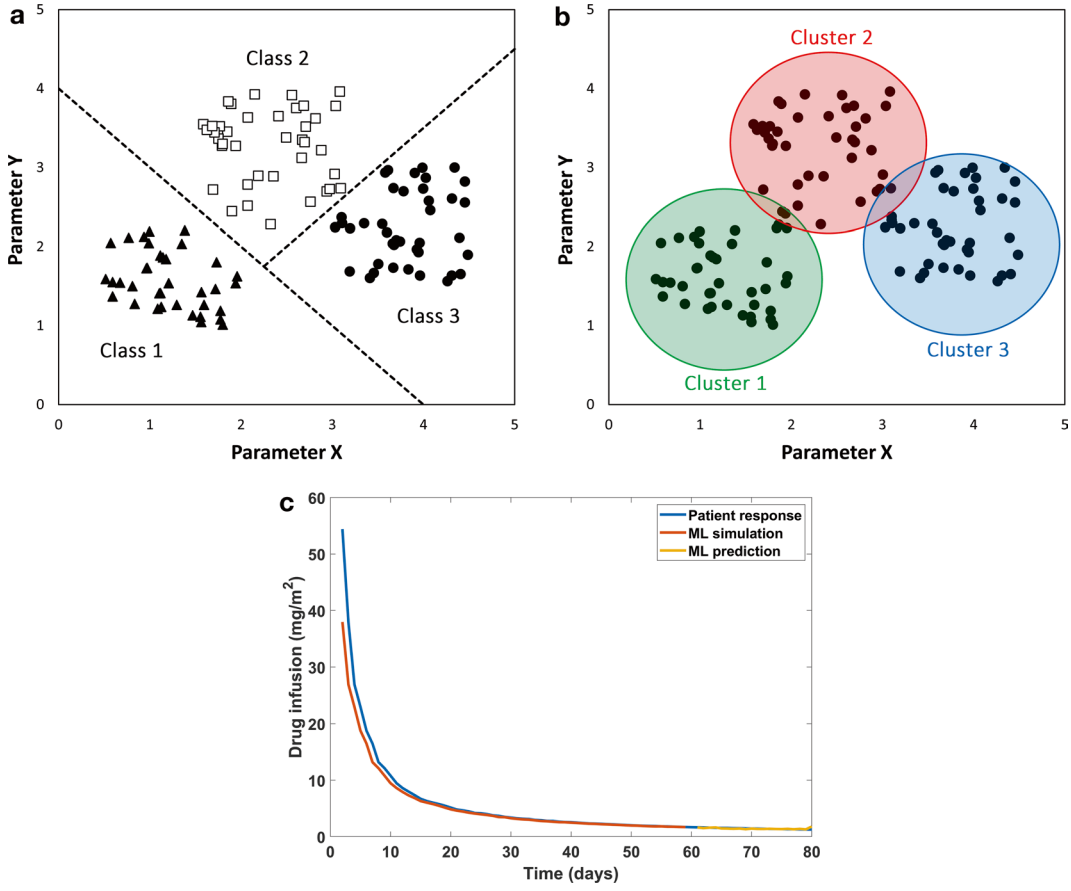

Fig. 2.2 Examples of (**a**) classification, (**b**) clustering and (**c**) prediction