Towards Trustworthy AI in Dentistry

J. Ma[1], L. Schneider[2,3] (iD), S. Lapuschkin[1], R. Achtibat[1], M. Duchrau[2], J. Krois[2,3] (iD), F. Schwendicke[2,3] (iD), and W. Samek[1,4] (iD)

Abstract

Medical and dental artificial intelligence (AI) require the trust of both users and recipients of the AI to enhance implementation, acceptability, reach, and maintenance. Standardization is one strategy to generate such trust, with quality standards pushing for improvements in AI and reliable quality in a number of attributes. In the present brief review, we summarize ongoing activities from research and standardization that contribute to the trustworthiness of medical and, specifically, dental AI and discuss the role of standardization and some of its key elements. Furthermore, we discuss how explainable AI methods can support the development of trustworthy AI models in dentistry. In particular, we demonstrate the practical benefits of using explainable AI on the use case of caries prediction on near-infrared light transillumination images.

Keywords: computer vision/convolutional neural networks, artificial intelligence, deep learning/machine learning, dental informatics/bioinformatics, mathematical modeling, standardization

Introduction

Oral and dental pathologies are among the most prevalent conditions of humankind, and direct costs for managing them as well as the associated indirect costs have been quantified at over 500 billion USD in 2015 (Righolt et al. 2018). Given demographic and epidemiological dynamics, this burden of oral and dental diseases is expected to grow, while the workforce to provide oral and dental care is limited, which in combination is stressing already strained health care systems and putting the affordability and accessibility of oral and dental care at risk (World Health Organization 2020). Digital technologies, such as artificial intelligence (AI), are frequently considered to make processes more efficient or increase the quality of decisions. Especially in oral and dental care, AI shows great promise due to the potential of higher effectiveness, safety, and efficiency, which allow provision of better care to a larger number of people (Schwendicke and Krois 2022). However, AI potentially introduces new risks that must be considered. Due to the design and complexity of AI algorithms, the outputs of such are often inexplicable, which limits the acceptance and hence the use of these methods. Explainable AI (XAI) is a rapidly evolving research field that is precisely concerned with finding solutions to encounter this effect (Holzinger et al. 2022). To guarantee a safe usage of AI solutions in health care and, in particular, in dentistry, different organizations are developing new standards and tools to ensure the safety, performance, and trustworthiness of AI solutions. Within this study, we present elements of AI standardization and demonstrate for a caries classification model how compliance with these standards can be supported through explainable AI to ensure trustworthiness of AI solutions within dentistry.

Standardization

The Role of Standardization and Standardization Activities

With the purpose of an AI-supported software that should be used globally, there are numerous aspects such as terminology, interoperability, safety, trustworthiness, risk management, and governance that are relevant for standardization. These and other aspects are, for instance, considered in international standardization organizations such as

- ISO: International Organization for Standards
- IEC: International Electrotechnical Commission
- ITU: International Telecommunication Union

Standardization organizations are often collaborating to maximize harmonization efforts. One prominent example of a very successful joint standardization effort is, for instance, the development of the High Efficiency Video Coding (HEVC)

[1]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin, Berlin, Germany
[3]ITU/WHO Focus Group on AI for Health, Topic Group Dental Diagnostics and Digital Dentistry, Geneva, Switzerland
[4]BIFOLD–Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

Corresponding Author:
L. Schneider, Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin, Aßmannshauser Str. 4-6, Berlin, 14197, Germany.
Email: schneider.lisa@charite.de

Table. Published Standards (Green) and Ongoing Standardization Projects (Blue).

AI principles ISO/IEC AWI TR 24372	Data quality ISO/IEC WD 5259-1	Risk management ISO/IEC CD 23894	AI terminology ISO/IEC CD 22989	Terminology of AI safety IEEE P2802
AI robustness ISO/IEC NP 24029	Assessment of AI systems ISO/IEC WD TS 4213	AI testing ETSI DTR INT 008 TR 103 821	Ethics ISO/IEC AWI TR 24368	. . .
AI principles ISO/IEC TR 24028	Data quality ISO/IEC 25012	Risk management ISO 14971	Medical devices IEC 60601	Functional safety ISO 61508
Software quality ISO/IEC 25000	Conformity assessment ISO/IEC 17000ff	Software life cycle ISO 62304	Big data ISO/IEC TR 20547	. . .

This overview is incomplete and only highlights some areas that are particularly relevant for this underlying work. The specified document number shows a standard/standardization project belonging to the respective topics. There are further documents/projects.
AI, artificial intelligence; IEC, International Electrotechnical Commission; ISO, International Organization for Standards. This table is available in color online.

standard that was developed by the Joint Video Experts Team (JVET) in a collaboration of the Video Coding Experts Group (VCEG) of ITU and the Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29.

The development of AI standardization is compared to some other standardization activities, such as video compression, still at an early stage. There exist standards and standards under development by the abovementioned organizations, for example, on the life cycle of medical device software (DIN 2016), the trustworthiness of AI (ISO 2020), risk management of medical devices (Deutsches Institut für Normung 2009), and many other important characteristics such as performance metrics of AI (ISO/IEC WD TS 4213; see also the Table).

However, there are still many challenges left, and there is no harmonized widely accepted standard for treatment of AI in health. This is partially due to the rapid development of the field of AI. ISO and IEC have a joint standardization committee—the ISO/IEC JTC 1/SC 42 Artificial Intelligence—that is developing standards in the general area of AI and acts as a focus point of AI standards development of the joint technical committee (JTC) 1. Many of its subgroups, in particular working groups (WGs), are working on aspects such as Data (WG2) and Trustworthiness (WG3). On the European standardization level there are, for instance,

- CEN: European Committee for Standardization
- CENELEC: European Committee for Electrotechnical Standardization
- ETSI: European Telecommunications Standards Institute

There are similar collaborations between these standardization organizations; for example, CEN and CENELEC have a joint standards committee, the CEN-CENELEC JTC 21 Artificial Intelligence, that itself has close collaborations with the ISO/IEC JTC 1/SC 42.

Next, we present a standardization effort by the ITU specifically dedicated to the standardization of AI in dentistry.

Standardization of AI in Dentistry

The ITU has several study groups that also cover different areas of AI. One focus group that runs under Study Group 16 is the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H). This focus group develops a standardized assessment framework of AI-based methods based on a number of different medical use cases. One of the use cases is Dental Diagnostics and Digital Dentistry, worked on by a Topic Group (TG-Dental), which already developed a checklist for authors, reviewers, and readers for AI in dental research (Schwendicke et al. 2021). The topic group was established in 2019 and considers itself a community of stakeholders from the medical and AI communities with a shared interest in the topic. At the time of writing, TG-Dental consists of 34 members from 18 countries and 5 continents. The members come from academia, industry, and the private sector and have defined different subtopics such as Operative and Cariology, Prosthodontics, Periodontal, Surgical, Oral Medicine and Maxillofacial Radiology, Endodontics, and Orthodontics, in each of which the members aim to establish processes and requirements to facilitate standardization and, specifically, benchmarking (i.e., standardized testing) of AI applications in dentistry.

Next we will explore some of the elements of AI standardization in more detail.

Elements of AI Standardization

There are numerous national documents from different countries discussing the needs of standards for AI and the challenges of developing such (e.g., National Institute of Standards and Technology 2019; Wahlster and Winterhalter 2020). The development of standards is a multidisciplinary effort and as AI tools are becoming more advanced and products are brought to market, legal and regulatory aspects are indispensable, adding another dimension that must be regarded. Recently, the European Union has published a proposal for a regulation to outline its perspectives on harmonized rules on AI (Artificial Intelligence Act 2021).

From a developing perspective and with a focus on trustworthiness of AI models, there is a clear demand for the following:

- *Data sets:* High-quality data sets are generally needed at different stages of an AI life cycle. They are particularly important during the development phase (e.g.,

during model training to create highly accurate and good performing AI models). However, high-quality data sets are also important in the evaluation of AI models, which are usually based on specific quality criteria and benchmarked using high-quality data sets against other methods.

- *Quality criteria:* Different quality criteria can be required to award satisfying competences to an AI method. These quality criteria may depend on its user and must be developed in a domain-specific manner. General technical quality attributions include performance, robustness, uncertainty quantification, and explainability. For data annotation, additional quality criteria are needed. The annotation process itself must be outlined in detail, as annotation requirements vary significantly depending on the actual task. Current gold standards are usually derived by majority voting schemes or external expert boards.
- *Tools:* To increase the common understanding, practicability, and usefulness of carefully designed quality attributions of an AI model, the development of tools that support the quality analysis should not be neglected. Such tools can be used to evaluate the model quality and should potentially be developed with the intention to be accessible and usable by a third party. The set of tools is not limited to evaluation of the model performance but can also provide insights (e.g., via explainability tools) into what strategies and representations the model has learned, as outlined below. Having well-developed and standardized evaluation tools could potentially increase the speed of product development and bring medical devices faster, yet still secure, to market.
- *Benchmarking:* The establishment of processes and ratings is helpful to rank and compare AI methods against each other. In particular, a comprehensive benchmarking system, including heterogeneous and representative test data, would increase the transparency of the model performance and allow end users to better compare different products.

It is important to stress that the trustworthiness of AI has a very large scope and deserves further research on securing the proper implementations of rules and guidelines. For example, the European Commission has outlined (European Commission and Directorate-General for Communications Networks, Content and Technology 2019) 7 key requirements: 1) human agency and oversight; 2) technical robustness and safety; 3) privacy and data governance; 4) transparency; 5) diversity, nondiscrimination, and fairness; 6) environmental and societal well-being; and 7) accountability. In this work, we primarily focus on technical work (including standards and XAI) that can be used to assist the implementation of trustworthy AI.

The development of technical standards for trustworthy AI models is ideally associated with tools that support the verification process and analysis of AI. XAI provides a fundamental access toward understanding what is usually considered a black box model. XAI can provide insights to performance metrics such as loss or accuracy ratings by exposing its internal representations and reasoning strategies. As such, XAI allows to establish more holistic and informed AI life cycles that seek to not only increase model performance but also fulfill predefined (e.g., clinical) requirements by optimizing the model behavior through feedback from domain experts.

Explainable AI in Dentistry

AI models in dentistry ideally follow equal patterns as dentists in their decision-making, while being able to execute them at higher speed and higher accuracy. Identifying deviating decision-making patterns allows to gauge possible bias, for example, by confounding or artifacts, and can help to increase the trust of users and recipients of AI. To demonstrate how such decision-making of AI may be analyzed via XAI to check compliance with developed AI standards, we consider in the following the prediction of caries lesions on individual teeth based on images obtained using near-infrared light transillumination (NILT) as 1 exemplary use case.

Data

The data set consists of 834 NILT images from routine examinations of 56 patients aged 18 y or older, recorded at Charité–Universitätsmedizin Berlin between 2019 and 2022 with ethical approval (EA4/080/18). Images were cropped around the central tooth and pixel-based annotations of caries lesions provided by 3 dental experts with clinical experience of 8 to 11 y under the standardized conditions of a custom-made annotation tool (Ekert et al. 2019). Annotations were revised by 1 master reviewer, who curated (reviewed, added, deleted) the annotations. Resulting segmentations were united and translated into binary class labels, resulting in 44% images containing caries lesions and 56% no lesions. Images were resized to a resolution of (224×224) pixels, and adaptive histogram equalization was performed for each image. Finally, the data were split into training, validation, and test sets with ratios of 80%, 10%, and 10%, respectively.

Model

The model is based on a slightly modified VGG-11 (Simonyan and Zisserman 2015) architecture, which was pretrained on ImageNet (Deng et al. 2009). Training with an augmented data set was performed over 200 epochs with the Adam Optimizer (learning rate: 10^{-4}). The training process was stopped when there was no improvement on the validation loss for 30 epochs. There was no extensive hyperparameter search performed, as we aimed to demonstrate AI reasoning processes instead of maximizing model performance.

Explaining Caries Predictions

In recent years, many approaches to explain predictions of AI models have been developed in the field of XAI (Samek et al. 2021). The subfield of local XAI assembles methods typically
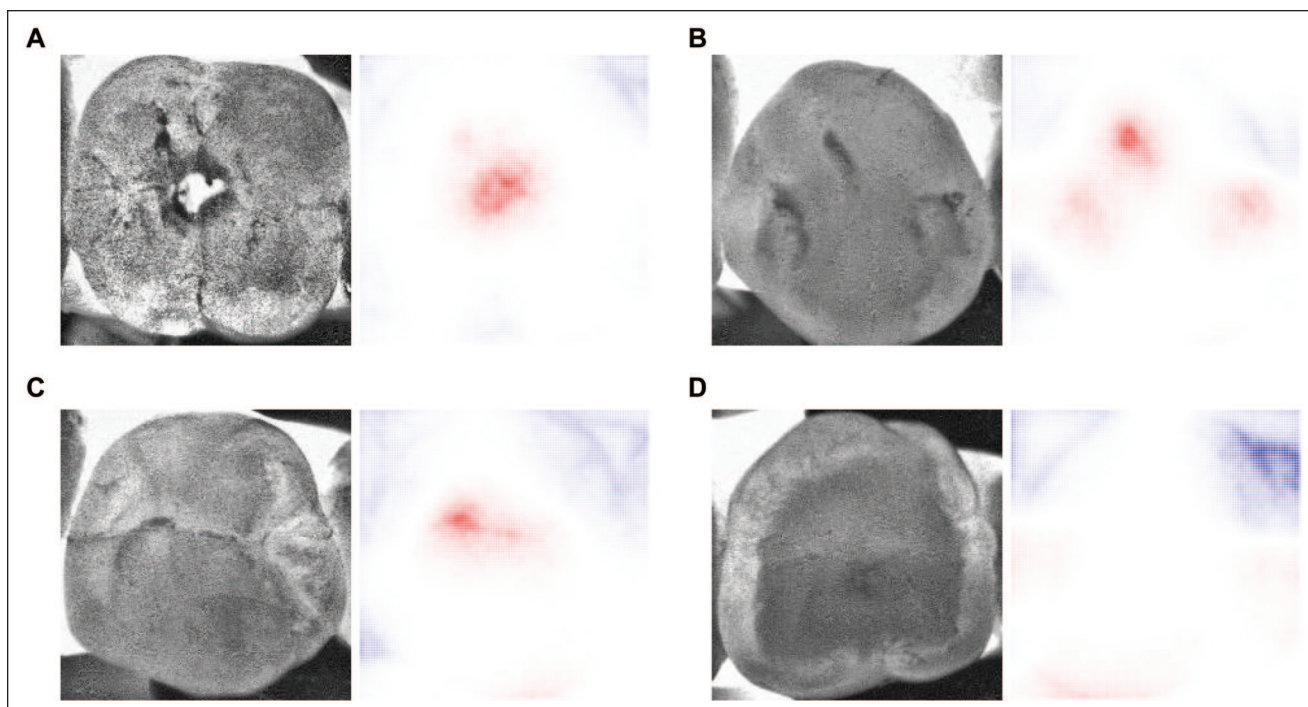
Figure. Correctly predicted near-infrared light transillumination images with corresponding layer-wise relevance propagation heatmaps, which point out areas in the image that the AI model considered as relevant for its decision-making process. The red-highlighted areas are in favor of the caries class, while blue areas show features relevant for the noncaries class. (**A**, **B**) Teeth of class caries with confidences of 0.97 and 0.98, respectively. (**C**, **D**) Noncaries class, which was correctly classified with confidences of 0.29 and 0.11. Reported confidence values reflect certainty of prediction for caries.

by computing attribution maps, that is, per-input dimension indicators of how important those given units are for the model during a particular inference on a specific sample.

Attribution maps can be obtained from various processes and carry different meanings, for example, how (much) a model has used the value of a particular input unit during inference (Bach et al. 2015; Shrikumar et al. 2017), or whether the model is sensitive to its change (Morch et al. 1995; Baehrens et al. 2010). This information can be obtained either via perturbation-based approaches treating the model as a black box (Zeiler and Fergus 2014; Ribeiro et al. 2016) (at the cost of increased computational cost), based on the gradient (Simonyan et al. 2014; Sundararajan et al. 2017) (given the prediction function is differentiable), or with techniques applying a modified backward pass through the model (Bach et al. 2015; Shrikumar et al. 2017). In this work, we employ the popular layer-wise relevance propagation (LRP) (Bach et al. 2015) using parameters recommended for the employed VGG-11 model (Kohlbrenner et al. 2020).

Results

The quality criteria of trustworthy AI models considered in AI standardization require clinically backed and explainable decisions. A dentist may perform the discussed caries classification task manually by scanning the tooth for abnormalities of darker gray shades with special attention to fissures and transitional areas to adjacent teeth. The characteristics of potential abnormalities allow dentists to distinguish between nondecayed and decayed teeth. This manual assessment is currently a standard procedure in dentistry. In future, this may be supported by highly efficient prescreening through AI, in which only difficult cases are assigned to dental experts. Such machine-aided diagnostics reduce time-consuming work, and the prediction of an AI model may be considered a second opinion, which is especially helpful for less experienced dentists. XAI may assist further by indicating areas that carried relevance for the model decision.

Figure A, B shows input images with their corresponding visualizations of LRP-based attribution maps of correctly classified samples with caries lesions. The explanations highlight that given the marked areas, the model decides to classify both teeth as decayed with high certainty with confidences of 0.97 and 0.98. Reported confidence values present the reliability of the predictions: 1.0 means the model is certain that the tooth shows a caries lesion and 0.0 stands for a noncaries tooth. The used decision threshold was 0.5. The highlighted areas carry also clinical relevance for caries detection by a dentist. Figure C, D shows correctly classified samples of teeth without caries lesions. Confidences of 0.24 and 0.11 indicate lower certainties than before. This uncertainty is reflected in the red-highlighted areas that argue for the caries class. Nonetheless, this behavior is reasonable, as the dentist would examine these areas as well during diagnostics.

The explanations further reveal that the model included the image corners for its decision-making. This behavior originates from the different perceptions and experiences of humans and AI models: the reasoning of dentists is based on general knowledge and lifelong training, through which they neglect background matters, such as the image corners, from examination. The AI model instead assumes that relevant information may be located anywhere in the image. Above that, it is only able to pick up its knowledge from the data it has been trained on. If the data contain confounding features that are statistically linked to a certain outcome, the model will pick up those features regardless of their clinical relevance. Those flaws may not affect the performance metrics and therefore remain uncovered. Nevertheless, such model behavior is clinically not reasonable and therefore should trigger model enhancements.

XAI tools can help to reveal such facet and may later be used in an extended feedback loop to observe the effect of performance-improving strategies, which may even directly fix the aspect within the model behavior (Anders et al. 2022).

Conclusion

In order to achieve trustworthiness of AI, experts from different communities and scientific fields have to work together and join forces. Only then meaningful and acceptable standards can be developed that secure the high quality of medical AI applications. Furthermore, we need high-quality tools such as XAI providing insights on individual predictions, but also the general reasoning of a model (e.g., via large-scale behavioral analyses) (Lapuschkin et al. 2019). In this manner, healthy AI life cycles can be established, leading to high-quality and representative data sets and models to constantly develop and improve the current state of research.

Author Contributions

J. Ma, contributed to conception, design and data interpretation, drafted and critically revised the manuscript; L. Schneider, S. Lapuschkin, contributed to conception, design, data analysis and interpretation, drafted and critically revised the manuscript; R. Achtibat, M. Duchrau, contributed to data analysis and interpretation, critically revised the manuscript; J. Krois, W. Samek, contributed to conception and design, critically revised the manuscript; F. Schwendicke, contributed to conception and design, drafted and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: F. Schwendicke and J. Krois are cofounders of the dentalXrai GmbH, a startup. dentalXrai GmbH did not have any role in conceiving, conducting, or reporting this study.

Funding

ORCID iDs

L. Schneider [iD] https://orcid.org/0000-0002-4431-2669

J. Krois [iD] https://orcid.org/0000-0002-6010-8940

F. Schwendicke [iD] https://orcid.org/0000-0003-1223-1669

W. Samek [iD] https://orcid.org/0000-0002-6283-3265

References

Anders CJ, Weber L, Neumann D, Samek W, Müller K-R, Lapuschkin S. 2022. Finding and removing Clever Hans: using explanation methods to debug and improve deep models. Inf Fusion. 77:261–295.

Artificial Intelligence Act. 2021. Proposal for a regulation of the European parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. EUR-Lex-52021PC0206 [accessed 2022 May 24]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. 10(7):1–46.

Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller KR. 2010. How to explain individual classification decisions. J Mach Learn Res. 11(2010):1803–1831.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. Imagenet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL. IEEE. p. 248–255. doi:10.1109/CVPR.2009.5206848

Deutsches Institut für Normung. Normenausschuss Medizin. 2009. Medizinprodukte: Anwendung des Risikomanagements auf Medizinprodukte (ISO 14971: 2007, korrigierte Fassung 2007-10-01); deutsche Fassung EN ISO 14971: 2009. Berlin, Germany: Beuth Verlag.

DIN, EN. 2016. 62304: 2016-10. Medizingeräte-Software-Software-Lebenszyklus-Prozesse (IEC 62304: 2006+ A1: 2015); Deutsche Fassung EN 62304: 2006+ Cor.: 2008+ A1: 2015.

Ekert T, Krois J, Meinhold L, Elhennawy K, Emara R, Golla T, Schwendicke F. 2019. Deep learning for the radiographic detection of apical lesions. J Endod. 45(7):917–922.e5.

European Commission and Directorate-General for Communications Networks, Content and Technology. 2019. Ethics guidelines for trustworthy AI. Luxembourg, Luxembourg: Publications Office of the European Union. doi:10.2759/346720

Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W. 2022. XxAI-beyong explainable artificial intelligence. Paper presented at: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. Berlin: Springer. p. 3–10 [accessed 2022 May 24]. https://link.springer.com/chapter/10.1007/978-3-031-04083-2_1

International Organization for Standardization (ISO). 2020. ISO/IEC TR 24028:2020 information technology—artificial intelligence—overview of trustworthiness in artificial intelligence [accessed 2022 May 24]. https://webstore.iec.ch/publication/67138

Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. 2020. Towards best practice in explaining neural network decisions with LRP. Paper presented at: 2020 International Joint Conference on Neural Networks (IJCNN). New York: IEEE. p. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206975

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun. 10(1):1096.

Morch NJ, Kjems U, Hansen LK, Svarer C, Law I, Lautrup B, Strother S, Rehm K. 1995. Visualization of neural networks using saliency maps. Paper presented at: Proceedings of ICNN. Volume 4. New York: IEEE. p. 2085–2090. https://doi.org/10.1109/ICNN.1995.488997

National Institute of Standards and Technology. 2019. U.S. leadership in AI: a plan for federal engagement in developing technical standards and related tools (prepared in response to Executive Order 13859) (2019 August 9). Washington (DC): US Department of Commerce.

Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. Paper presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 2016. New York, (NY): Association for Computing Machinery. p. 1135–1144. https://doi.org/10.1145/2939672.2939778

Righolt A, Jevdjevic M, Marcenes W, Listl S. 2018. Global-, regional-, and country-level economic impacts of dental diseases in 2015. J Dent Res. 97(5):501–507.

Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R. 2021. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE. 109:247–278 [accessed 2022 May 24]. https://ieeexplore.ieee.org/document/9369420

Schwendicke F, Krois J. 2022. Data dentistry: how data are changing clinical care and research. J Dent Res. 101(1):21–29.

Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J. 2021. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent 107:103610.

Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. Paper presented at: Proceedings of the 34th International Conference on Machine Learning. Vol. 70. Sydney, NSW, Australia. p. 3145–3153.

Simonyan K, Vedaldi A, Zisserman A. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. International Conference on Learning Representations ICLR 14. abs/1312.6034 [accessed 2022 May 24]. https://openreview.net/forum?id=cO4ycnpqxKcS9

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. ARVIX PREPRINT [accessed 2022 May 24]. https://arxiv.org/abs/1409.1556

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. Paper presented at: Proceedings of ICML. ICML'17. JMLR.org. p. 3319–3328. ARVIX PREPRINT [accessed 2022 May 24]. http://www.arxiv-vanity.com/papers/1703.01365

Wahlster W, Winterhalter C. 2020. Deutsche Normungsroadmap Künstliche Intelligenz. DIN & DKE 11. Berlin, Germany: DIN e.V. and DKE.

World Health Organization. 2020. Achieving better oral health as part of the universal health coverage and noncommunicable disease agendas towards 2030. Report to Director General 2020. Geneva (Switzerland): WHO.

Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. Paper presented at: Proceedings of European Conference on Computer Vision. Vol. 8689 of Lecture Notes in Computer Science. Berlin: Springer. p. 818–833.