Towards trustworthy AI in dentistry

Jackie Ma[1], Lisa Schneider[2,3], Sebastian Lapuschkin[1], Reduan Achtibat[1], Martha Duchrau[2], Joachim Krois[2,3], Falk Schwendicke[2,3], Wojciech Samek[1,4]

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, 1211 Geneva, Switzerland
4 BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

Short title: Trustworthy AI in dentistry

Keywords: Computer Vision/Convolutional Neural Networks, Artificial Intelligence, Deep Learning/Machine Learning, Dental informatics/bioinformatics, Mathematical modelling, Standardization

Corresponding author:

M.Sc. Lisa Schneider
Charité – Universitätsmedizin Berlin
Department of Oral Diagnostics, Digital Health and Health Services Research, Charité - Universitätsmedizin Berlin, Germany
Aßmannshauser Str. 4-6
14197 Berlin, Germany
schneider.lisa@charite.de

Word count: -
Number of tables: 1
Number of figures: 1
Number of references: 26

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. Further, 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.

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 healthcare 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 it's the acceptance and hence the use 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 healthcare 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, governance, etc. 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) 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 above-mentioned organization, e.g. 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 Table 1.

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 (WG) 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 as 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 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

- *Datasets*: High quality datasets 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 a highly accurate and good performing AI models. However, high quality data

sets are also important in the evaluation of AI models, which is usually based on specific quality criteria and benchmarked using high quality data sets against other methods.

- *Quality criteria*: There are different quality criteria that 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 varies 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 evaluate the model performance but can also provide insights, e.g. via explainability tools 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 are 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) seven key requirements: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination 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 are ideally associated with tools that support the verification process and analysis of AI. Explainable AI (XAI) provides a fundamental access towards 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 which seek to not only increase model performance, but 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 artefacts, 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 one exemplary use case.

Data

The dataset consists of 834 NILT images from routine examinations of 56 patients aged 18 years or older, recorded at Charité - Universitätsmedizin Berlin between 2019-2022 with ethical approval (EA4/080/18). Images were cropped around the central tooth and pixel-based annotations of caries lesions provided by three dental experts with clinical experience of 8-11 years under the standardized conditions of a custom-made annotation tool (Ekert et al. 2019). Annotations were revised by one 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 was split into training, validation and test set 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 pre-trained 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 sub-field of local XAI assembles methods typically 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, e.g., 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 non-decayed and decayed teeth. This manual assessment is currently a standard procedure in dentistry. In future, this may be supported by highly efficient pre-screenings 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 as a second opinion, which is especially helpful for less experienced dentists. XAI may assist further by indicating areas which carried relevance for the model decision.

Figures 1a and 1b show 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 non-caries tooth. The used decision threshold was 0.5. The highlighted areas carry also clinical relevance for caries detection by a dentist. Figures 1c and 1d show correctly classified samples of teeth without caries lesions. Confidences of 0.24 and 0.11 indicate lower certainties as 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 life-long 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 contains confounding features which 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 utilized 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. Further, 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 analyzes (Lapuschkin et al. 2019). In this manner healthy AI life cycles can be established, leading to high quality and representative datasets and models, to constantly develop and improve the current state of research.

Acknowledgement

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) through the Berlin Institute for the Foundations of Learning and Data (BIFOLD) under Grant 01IS18025A and Grant 01IS18037I. Further, this work was funded by Deutsche Forschungsgemeinschaft (DFG) under Grant KR 5457/1. Above that, this study did not receive any support from outside and there is nothing to acknowledge.

Conflict of interest

FS and JK are co-founders of the dentalXrai GmbH, a startup. dentalXrai GmbH did not have any role in conceiving, conducting or reporting this study.

Author's contribution

JM, LS, SL, JK, FS, WS conceived and designed the study. JM, LS, SL, FS wrote the manuscript. LS, SL, RA conducted the study. JM, LS, SL, RA, MD, FS analyzed the data. All authors revised the paper and gave their final approval and agree to be accountable for all aspects of the work.

References

Act AI. 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.

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. Information Fusion 77:261–295. doi:https://doi.org/ 10.1016/j.inffus.2021.07.015.

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 K-R. 2010. How to explain individual classification decisions. JMLR 11:1803–1831.

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

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. Journal of Endodontics 45(7). doi:10.1016/j.joen.2019.03.016.

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

International Organization for Standardization [ISO]. 2020. ISO/IEC TR 24028:2020 information technology – artificial intelligence – overview of trustworthiness in artificial intelligence.

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. Beuth

Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W. 2022. Xxai-beyong explainable artificial intelligence. In: International Workshop on Extending Exlainable AI Beyond Deep Models and Classifiers. Springer. P.3-10

Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. 2020. Towards best practice in explaining neural network decisions with lrp. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE. p. 1–7. doi: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. Nature communications 10:1–8.

Morch NJ, Kjems U, Hansen LK, Svarer C, Law I, Lautrup B, Strother S, Rehm K. 1995. Visualization of neural networks using saliency maps. In: Proc. ICNN. volume 4. IEEE. p. 2085–2090.

National Institute of Standards and Technology. 2019. US Leadership in AI: A plan for Federal Engagement in Developing Technical Standards and Related Tools. US Department of Commerce Washington, DC.

Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?" explaining the predictions of any classifier. In: Proc. ACM SIGKDD. p. 1135–1144.

Righolt A, Jevdjevic M, Marcenes W, Listl S. 2018. Global-, regional-, and country-level economic impacts of dental diseases in 2015. Journal of dental research 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. Proceedings of the IEEE 109:247–278.

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. Journal of Dentistry 107:103610. doi:https://doi.org/10. 1016/j.jdent.2021.103610.

Schwendicke F, and Krois J. 2022. Data Dentistry: How Data Are Changing Clinical Care and Research. Journal of Dental Research 101(1):21–29.Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. In: Proc. ICML. volume 70 of Proc. MLR. PMLR. p. 3145–3153.

Simonyan K, Vedaldi A, Zisserman A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. ICLR. volume abs/1312.6034.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for largescale image recognition. 1409.1556.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. In: Proc. ICML. ICML'17. JMLR.org. p. 3319–3328.

Wahlster W, Winterhalter C. 2020. Deutsche Normungsroadmap Künstliche Intelligenz. DIN & DKE 11.

World Health Organization, Oral health. 2020. Achieving better oral health as part of the universal health coverage and noncommunicable disease agendas towards 2030, in: Report to Director. General 2020.

Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. In: Proc. ECCV. volume 8689 of Lecture Notes in Computer Science. Springer. p. 818–833.

Tables

Table 1: List of published standards (blue) and ongoing standardization projects (green). Note this overview is incomplete and only highlights some areas that 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 Principles (ISO/IEC TR 24028)	Software Life Cycle (ISO 62304)	Conformity Assessment (ISO/IEC 17000ff)	Big Data (ISO/IEC TR 20547)	Software Quality (ISO/IEC 25000)
Data Quality (ISO/IEC 25012)	Risk Management (ISO 14971)	Functional Safety (ISO 61508)	Medical Devices (IEC 60601)	…
AI principles (ISO/IEC AWI TR 24372)	AI terminology (ISO/IEC CD 22989)	Data Quality (ISO/IEC WD 5259-1)	Risk Management (ISO/IEC CD 23894)	AI testing (ETSI DTR INT 008 (TR 103 821))
Ethics (ISO/IEC AWI TR 24368)	Terminology of AI safety (IEEE P2802)	Assessment of AI Systems (ISO/IEC WD TS 4213)	AI robustness (ISO/IEC NP 24029)	…

Figure Captions

Figure 1: Correctly predicted NILT images with corresponding LRP heatmaps, which point out areas in the image that the AI model considered as relevant for its decision-making process. The red color highlights areas which spoke in favor of the caries class, while blue areas show features relevant for the non-caries class. (a) and (b) show teeth of class caries with confidences of 0.97

and 0.98, respectively. The non-caries class is represented in (c) and (d), which were correctly classified with confidences of 0.29 and 0.11. Reported confidence values reflect certainty of prediction for caries.