SkinSplain: An XAI Framework for Trust Calibration in Skin Lesion Analysis

Tim Katzke[1,2,*,†], Mustafa Yalçıner[1,2,†], Jan Corazza[1,2,†], Alfio Ventura[1,3,†], Tim-Moritz Bündert[4,**,†] and Emmanuel Müller[1,2]

[1]Research Center Trustworthy Data Science and Security, University Alliance Ruhr, Joseph-von-Fraunhofer-Str. 25, Dortmund, 44227, Germany

[2]TU Dortmund University, August-Schmidt-Straße 1, Dortmund, 44227, Germany

[3]University of Duisburg-Essen, Bismarckstraße 120, Duisburg, 47057, Germany

[4]Scientific Computing Center, Karlsruhe Institute of Technology, Zirkel 2, Karlsruhe, 76131, Germany

Abstract

Explainable artificial intelligence (XAI) methods provide insights into machine learning models by making their decision processes more transparent for humans. Ideally, such transparency enables users to trust the AI to an appropriate extent, with understanding both its capabilities and limitations for the given task. However, evaluations of XAI methods rarely assess their impact on users' perceived trust alignment with actual model capabilities. In fact, a recent survey reveals that 80% of published work introducing an XAI method does not include user studies. To bridge this gap, we introduce SkinSplain, a web-based framework designed for measuring users' perceived trust in AI systems when interacting with both numerical and visual interpretability cues. SkinSplain allows users to provide inputs to a machine learning model and observe its explanations for predictions. Crucially, users then self-report their level of trust in the model's predictions. These trust scores facilitate further analysis in user studies. Given the increased popularity of AI-based skin lesion analyzers, we employ SkinSplain in a user study to examine how explanation methods influence trust in AI-driven medical diagnostics. The source code is available at https://github.com/Ti-Kat/SkinSplain.

Keywords

XAI for Medical Diagnosis, Skin Lesion Analysis, Trust Calibration, Human-AI Interaction, Computer Vision

1. Introduction

It is challenging to understand and explain the factors contributing to an AI systems prediction. In a recent example, Winkler et al. found that markings by standard surgical ink markers in skin lesion images lead to significant changes in AI predictions [1]. That is due to biases in the training data, where some artefacts of medical doctors in the images correlate with high risks of the skin lesion being malignant. Such biases in the training data can cause the AI system to fail when employed on real-world data.

Trust is widely recognised as a central variable explaining user's resistance or over-reliance in automated systems [2]. Here, we consider trust from a strictly psychological perspective [3], derived from interpersonal trust [4], which was translated into trust in automation [2] and therefore synthetic relationships [5]. Further, trust calibration in AI refers to the alignment between a user's trust in an AI system and the system's actual capabilities. It involves preventing overtrust, where users blindly use an incapable AI system, and undertrust, where users refuse to use the AI system, despite the system's

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.

**Work done while at Research Center Trustworthy Data Science and Security.

†These authors contributed equally.

tim.katzke@tu-dortmund.de (T. Katzke); mustafa.yalciner@tu-dortmund.de (M. Yalçıner); jan.corazza@tu-dortmund.de (J. Corazza); alfio.ventura@uni-due.de (A. Ventura); tim-moritz.buendert@kit.edu (T. Bündert)

https://corazza.github.io/ (J. Corazza); https://rc-trust.ai/about/scientists/alfio-ventura (A. Ventura)

0009-0000-0154-7735 (T. Katzke); 0009-0005-6240-7062 (M. Yalçıner); 0009-0000-1342-0117 (J. Corazza); 0000-0003-1639-8001 (A. Ventura); 0009-0000-7228-6106 (T. Bündert); 0000-0002-5409-6875 (E. Müller)

reliability [6]. Proper trust calibration ensures users' confidence matches the AI's performance, leading to more effective decision-making and collaboration [6]. Accordingly, we understand trust-calibration as an alignment problem between subjective trust perception of the human and objective trustworthiness of the technical system [6, 2, 7].

Demonstrating to AI system users that they can control and evaluate the quality of input data on which AI predictions are performed can significantly enhance their ability to calibrate trust towards such systems [6, 8]. For example, allowing users to interact with the AI system and explore its behaviour on various inputs can help users build a nuanced understanding of when AI predictions are trustworthy and when human oversight is required. This, in turn, allows them to discern how manipulations to input data influence and may enhance the performance and trustworthiness of AI predictions.

However, this type of controllability is underexplored in the literature focusing on Explainable AI and trust in AI. In fact, a recent survey highlights that only one in five papers proposing a new XAI method conducts any form of user survey [9]. This highlights two problems in the current XAI research. First, new explainability methods are usually not evaluated on real user studies. Therefore, it remains unclear, which XAI method actually help calibrate users's trust in the AI model. Secondly, it remains underexplored how a user's ability to interact with the AI and select inputs for which the system performs well or poorly impacts the trust calibration. More specifically, recent reviews on trust calibration [6], trust in AI [10, 11, 12] as well as the "unified and practical user-centric framework for explainable artificial intelligence" [13], and experimental studies [14] do not address how explaining the role of human-controlled inputs in enhancing the performance of technical systems may help develop calibrated trust. Therefore, current literature falls short in guiding users of technical systems on how to achieve a collaborative performance that exceeds the capabilities of either humans or technical systems operating independently.

To close this gap, we design the web application SkinSplain. SkinSplain is a framework designed for aiding the explainability of an AI system by allowing the user to explore the system's behaviour on various inputs and understand model behaviour with the visual explainers. This interactive approach enables the user to investigate the nuances in the model's predictive performance, while being supported with explainers that facilitate model understanding. Crucially, the user can then report a trust score for each of the inputs, allowing for a subsequent analysis in a broader user study.

We demonstrate the use of SkinSplain practically and employ our framework in a preregistered study on skin cancer detection [1]. More specifically, we investigate whether (1) explaining how the AI model generates predictions and (2) showing inputs for which the AI systems' predictive performance deteriorates, leads to more calibrated trust in the AI system among laypeople.

2. Related Work

Whether layperson or expert, anyone who wants to evaluate new information with an established AI system must provide input data. Such highly interactive AI systems, which depend on high-quality user-controlled input, are relatively new. For example, large language models like ChatGPT and image-based applications like Foodvisor [2] are used in everyday life, and generate outputs based on whatever input is given. We were specifically inspired by skin cancer detection and prevention due to its high practical relevance [15, 16], current developments [17, 18, 19] and strong data availability [20]. SkinVision [3] and FotoFinder [4] are commercially available, clinically validated, regulated and certified applications that provide an AI prediction based on a photo of a skin lesion. This is done, for example, by automating the quantification of the ABCD rule [21] (Asymmetry, Border, Color, Diameter), by displaying the image areas that are particularly important for the prediction, or by assigning a score that indicates the overall risk.

[1] https://aspredicted.org/2ryf-7y88.pdf
[2] https://www.foodvisor.io/en/
[3] https://www.skinvision.com/
[4] https://www.fotofinder.de/en/

Currently, research on human controllability in such systems remains limited compared to more rigid, less interactive AI systems with low human-controllable aspects —as seen, for example, in AI-assisted decision-making [22, 23, 24]. A user's perceived control is essential in determining the user's intention to use a technical system [8], a stepping stone to actual usage experience [25, 26]. Thus, perceived control is also essential for optimal trust calibration that results from extensive usage experience. Theoretical considerations on trust in automation [2] emphasize that understanding the functionalities, strengths, weaknesses, and limitations of technical systems is crucial for trust calibration. This leads to an understanding when the technical system should and should not be used [7]. However, an understanding of the limitations —an understanding of the human-controllable elements of prediction quality— could lead to behavior fostering improved performance and trustworthiness in all situations and go beyond the general decision of usage. In fact, a series of studies [8] indicate that participants were more willing to use an imperfect algorithm if they could control it slightly. To summarize, emphasizing user controllability ensures that users obtain necessary learning experiences for long-term trust calibration [8, 25, 26], ultimately fostering an appropriate level of trust in technical systems over time [2].

3. The SkinSplain Framework

SkinSplain is a web-based framework that delivers real-time, interactive explanations of a classifier's decisions, enabling users to select and manipulate inputs while observing the immediate impact on both model output and explanation quality. SkinSplain integrates mechanisms for participants to provide self-reported trust measures directly within the interface. These perceived trust values may serve as valuable ground truth for assessing user confidence, allowing researchers to compare subjective evaluations with objective trust metrics, such as model predictive performance. As an application domain, SkinSplain focuses on skin lesion classification, where images are categorized as benign or malignant. This application not only underscores the practical relevance of the framework but also highlights the importance of aligning explainability with both user trust and empirical performance measures to ensure reliable and interpretable AI systems.

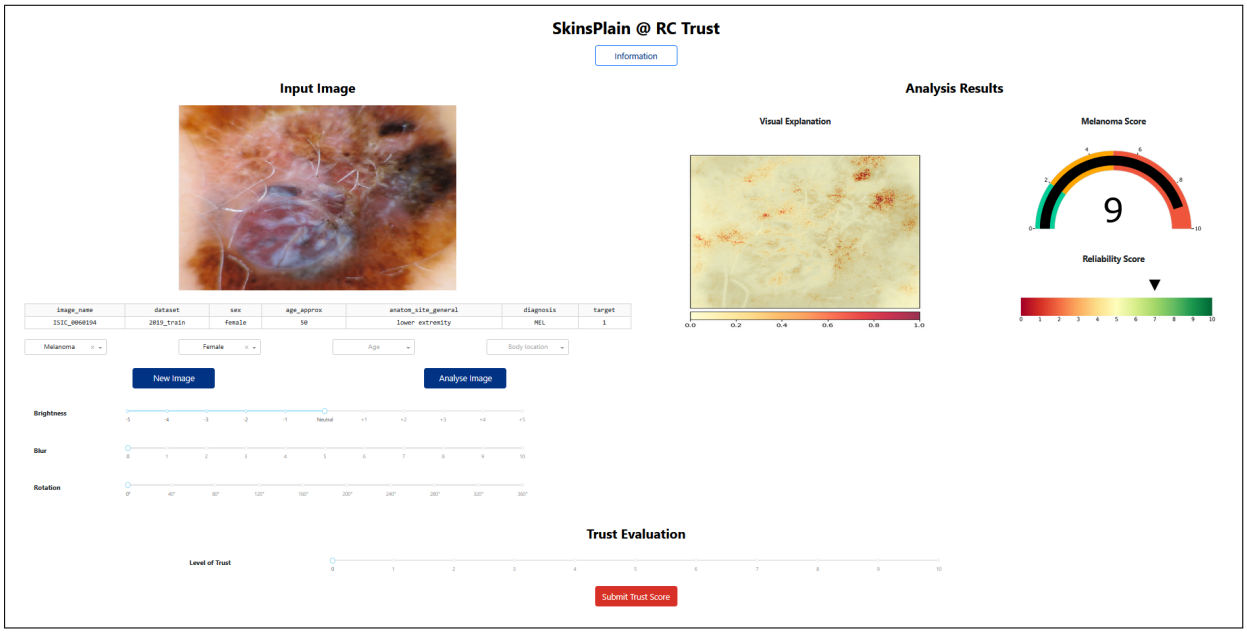


Figure 1: An overview of the SkinsPlain user interface. It consists of an area for image selection and augmentation by the user (left side), real-time analysis results of the selected image based on XAI methodologies (right side) and an input option for the user trust in these analysis results (bottom).

3.1. User Interface

The SkinSplain user interface, as shown in Figure 1, is divided into three sections.

Left Side (User-Controlled Input Selection) The left side of the interface provides users with direct control over the input selection, while also displaying the current input image along with additional metadata below. Users may load a new image from the ISIC skin lesion dataset [20] by clicking the "New Image" button. Prior to selection, they can apply filters based on demographic attributes. Multiple available drop-down menus correspond to categorical filters, such as age, diagnosis, sex, or lesion location. For instance, selecting the "Body location" filter reveals options like "torso" or "hand". To emulate realistic variations in image quality, users can also adjust brightness, blur, and rotation via interactive sliders, with changes immediately reflected on the screen. This functionality not only ensures that the input data covers varying real-world conditions, but also allows the users to have more influence over the input characteristics — a crucial factor in calibrating trust.

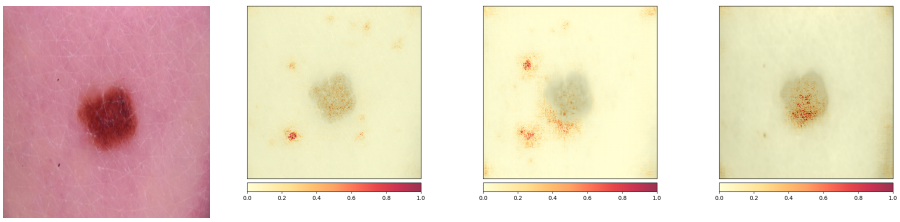
Still, to ensure ethical and responsible use, SkinsPlain is limited to publicly available ISIC data and does not support user-uploaded images. This restriction helps mitigate privacy risks and ethical concerns associated with applying an uncertified AI system to real user-provided medical images [27].

Right Side (XAI Analysis) Once the user clicks "Analyse Image", the results of one more more XAI methods are displayed on the right side of the interface within seconds, providing transparent communication of the model's internal decision-making processes. This transparency is essential for users to assess how self-controlled input adjustments affect model predictions. We briefly outline the currently integrated XAI methods below, and give a motivation and detailed explanations in Section 3.2.

- **Melanoma Score** – Prediction for the current input as benign or malignant on a scale of 1 to 10.
- **Reliability Score** – Assesses the reliability of the model's prediction for the given input.
- **Visual Explanation** – Highlights image regions most significant on the prediction.
- **Similar Images** – Displays the most similar represented images from both classes.

Bottom (Measuring Users Perceived Trust) After reviewing the XAI analysis results, users can submit their perceived trust in the model's prediction via a slider at the bottom. The compilation and evaluation of these perceived trust scores provide the necessary user trust self-report for systematic user trust calibration analyses. For the demo, we opted for a simple, one-dimensional measurement of trust perception. The basic SkinsPlain framework is designed for repeated presentation of inputs. We recommended limiting self-report measures within the framework to avoid overburdening participants and keep them motivated. These measurements become valid because of the repeated measures design. We recommend incorporating the SkinsPlain framework into a larger survey akin to our preregistered study if complex state and trait self-report measures are pursued.

3.2. AI Model and XAI Technologies



(a) Original image (b) Baseline (c) Blur + Brightened (d) Blur + Darkened

Figure 2: Comparison of a benign skin lesion (a) with visual explanations via Integrated Gradients saliency maps for the original image (b) and after applying gaussian blur with increased (c) and lowered (d) brightness.

This section outlines the AI model and associated XAI techniques employed in our framework.

AI Model Foundation: Skin Lesion Classification Given that our image data is based on the ISIC Challenge datasets, we referenced the winning solution from the 2020 ISIC Challenge [28] in developing our skin lesion classifier AI model. The solution employed an ensemble of convolutional neural networks (CNNs), based predominantly on the EfficientNet architecture. Since individual models in the ensemble performed nearly as well as the full ensemble, we opted for a simpler, single-model approach, based one a more recent variant of that architecture, to facilitate the application of standard XAI methods. Specifically, we fine-tuned an EfficientNetV2-S model [29] , pre-trained on ImageNet, for binary classification using images from the "Nevus" and "Melanoma" classes from a subset of the ISIC datasets, while deliberately excluding metadata such as demographic attributes. Our skin lesion classifier achieves an AUC-ROC of 0.9548 on an unseen test set drawn from the 2018 ISIC data.

Numerical AI-Trustworthiness Metrics: Melanoma and Reliability Scores The *Melanoma Score* offers an intuitive measure of the classifier's confidence, where a value of 0 indicates high confidence in benign classification and 10 indicates high confidence in malignancy. This is based on the classifier's single-neuron output, passed through a sigmoid activation function. The displayed melanoma score is obtained by linearly mapping the logit to a scale ranging from 0 to 10. Quantified on the same scale, the *Reliability Score* indicates how closely an input image aligns with the training data distribution. In essence, the further an input deviates from what the model is accustomed to, the less reliable its predictions become. Providing users with a quantitative measure of this reliability is crucial for informed decision-making. To that end, we calculate these scores for each input using a layer-wise variant of the *Deep k-Nearest Neighbors* [30] algorithm. This identifies the k-nearest neighbors from the training set within each layer of the skin lesion classifier, and computes a score that quantifies the consistency of the latent behavior of the input as it is processed by the model. By reducing complex model outputs to simple numerical indicators indicating objective trustworthiness— consistent in scale with the selectable levels of subjective perceived trust — these scores facilitate trust calibration, and enable users to quickly gauge the certainty and reliability of the current prediction. Long term, trust calibration for the given AI model and domain (here skin cancer detection) emerges, which may be transferred to similar technology.

Visual Interpretability: Saliency Maps and Similar Images *Saliency Maps* serve as a visual tool to highlight the most influential features that affect the classifier's decision. These influential features are often determined by analyzing how changes in the input impact the model's output [31]. We employ the Integrated Gradients method [32] on a transparent version of the base image for its robust estimation of feature importance, while also applying Gaussian noise to the output image to further smooth the results. This method is included to enhance interpretability by transparently communicating which input regions drive the classifier's output, and how this may change under diverse image manipulations. An example of this is visualized in Figure 2. Here it can be observed, that by increasing or decreasing the brightness, the model focuses either more or less on irrelevant image artifacts outside the actual skin lesion area. To communicate the behavior of the skin lesion classifier based on another visual cue, we also optionally display *Similar Images* from the training dataset with respect to the internally learned representation of the currently analysed image. This is performed for both a true melanoma and a true benign image. More precisely, these most similar images are determined by identifying the single nearest neighbors of either class in the representation space of the classifiers penultimate layer based on eudclidean distance. This motivated by the goal of highlighting similarities and distinctions between supposedly similar images with drastically different implications in a high-stakes environment.

4. Trust Calibration with SkinSplain

We now briefly outline potential research applications of SkinSplain for user trust calibration.

Evaluating XAI through User Interaction SkinSplain enables real-time, interactive exploration of model predictions alongside their corresponding explainability outputs, allowing study participants to investigate how input quality influences outcomes and to understand their role in a collaborative prediction process. Adding survey questions to the XAI-setup allows for repeated assessment of participant perceptions, facilitating interdisciplinary research, especially on trust calibration with its deep perceptional-technical nature [6, 2, 7]. Moreover, SkinSplain allows for the evaluation of various XAI methods, whether using a subset of the provided methods, or substituting alternative approaches, to systematically assess their impact on perceived user trust. When interactivity is not essential, the interface can be configured for fixed, survey-based studies (as shown in Figure 3 and advertised in our preregistration), offering a controlled environment for online experiments with XAI.

Balancing Perceived Trust and Objective Trustworthiness In our framework, user trust is shaped both by interactive, controllable inputs and by the interpretability cues presented. Trust calibration involves balancing this trust—that is, how much perceived user trust *is* attributed to the AI—with objective measures of the system's performance (e.g., accuracy or reliability) that indicate the trustworthiness of the system and how much trust *should* be placed into it. SkinSplain supports assessing both measures; self-reported trust measures can be obtained in flexible user studies that investigate the role of human controllable inputs and the influence of XAI methods on trust in conjunction with objective model performance. This provides the necessary data to systematically analyse trust calibration [6, 2, 7].

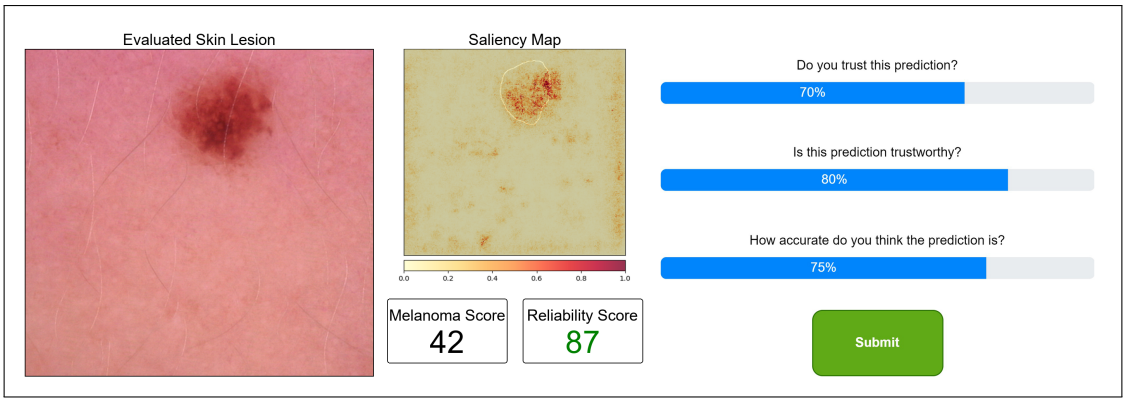


Figure 3: Example stimuli for a static case study, along with potential questions to evaluate user trust.

5. Discussion and Outlook

We introduced SkinSplain, a web application framework designed to investigate how XAI methods affect user behavior and trust perceptions. Our work demonstrates its use in user studies examining trust in a skin cancer classifier, where participants evaluated visual explainers and reliability measures. Although our current implementation targets the skin cancer domain, our framework is inherently domain-agnostic. Moreover, its components can be easily replaced to support user studies across diverse subsets of XAI methods. Looking ahead, we plan to extend our research with non-static user studies that exploit SkinSplain's interactive capabilities to further explore the influence of XAI on user behavior, ultimately contributing to the development of more trustworthy AI systems.

Acknowledgments

This research is funded by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centres within the University Alliance Ruhr (https://uaruhr.de).

Declaration on Generative AI

During the preparation of this proposal, we used the ChatGPT 4o model from OpenAI for minor language edits, aiming to enhance readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the manuscript's content.

References

[1] J. K. Winkler, K. Sies, C. Fink, F. Toberer, A. Enk, M. S. Abassi, T. Fuchs, H. A. Haenssle, Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition, European Journal of Cancer 145 (2021) 146–154.

[2] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors: The Journal of the Human Factors and Ergonomics Society 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.

[3] A. P. Association, trust, 2018. URL: https://dictionary.apa.org/trust.

[4] D. J. McAllister, Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations, Academy of Management Journal 38 (1995) 24–59. doi:10.5465/256727.

[5] C. Starke, A. Ventura, C. Bersch, M. Cha, C. de Vreese, P. Doebler, M. Dong, N. Krämer, M. Leib, J. Peter, L. Schäfer, I. Soraperra, J. Szczuka, E. Tuchtfeld, R. Wald, N. Köbis, Risks and protective measures for synthetic relationships, Nature Human Behaviour 8 (2024) 1834–1836. URL: https://www.nature.com/articles/s41562-024-02005-4. doi:10.1038/s41562-024-02005-4.

[6] M. Wischnewski, N. Krämer, E. Müller, Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ACM, 2023, pp. 1–16. doi:10.1145/3544548.3581197.

[7] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human Factors: The Journal of the Human Factors and Ergonomics Society 39 (1997) 230–253. doi:10.1518/001872097778543886.

[8] B. J. Dietvorst, J. P. Simmons, C. Massey, Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them, Management Science 64 (2018) 1155–1170. doi:10.1287/mnsc.2016.2643.

[9] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3583558. doi:10.1145/3583558.

[10] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021, pp. 624–635. doi:10.1145/3442188.3445923.

[11] A. D. Kaplan, T. T. Kessler, J. C. Brill, P. A. Hancock, Trust in artificial intelligence: Meta-analytic findings, Human Factors: The Journal of the Human Factors and Ergonomics Society 65 (2023) 337–359. doi:10.1177/00187208211013988.

[12] E. Glikson, A. W. Woolley, Human trust in artificial intelligence: Review of empirical research, Academy of Management Annals 14 (2020) 627–660. doi:10.5465/annals.2018.0057.

[13] S. Kaplan, H. Uusitalo, L. Lensu, A unified and practical user-centric framework for explainable artificial intelligence, Knowledge-Based Systems 283 (2024) 111107. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705123008572. doi:10.1016/j.knosys.2023.111107.

[14] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, M. Mara, Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task, Computers in Human Behavior 139 (2023). doi:10.1016/j.chb.2022.107539.

[15] H. G. V. Research, Europe ai in healthcare market size & outlook, 2023-2030, 2023. URL: https://www.grandviewresearch.com/horizon/outlook/ai-in-healthcare-market/europe.

[16] E. Commission, C. Directorate-General for Communications Networks, Technology, PwC, Study on eHealth, interoperability of health data and artificial intelligence for health and care in the European Union – Final study report. Lot 2, Artificial Intelligence for health and care in the EU, Publications Office of the European Union, 2021. doi:doi/10.2759/506595.

[17] O. Johnson, "World's first" AI skin cancer detection approved in Europe, Med-Tech Insights (2025). URL: https://med-techinsights.com/2025/02/06/worlds-first-ai-skin-cancer-detection-approved-in-europe/.

[18] U. Sahin, Özlem Türeci, Medicine, technology and the end of cancer - artificial intelligence may be the key to unlocking personalized cancer vaccines., The New York Times (2023). URL: https://www.nytimes.com/2023/12/05/special-series/artificial-intelligence-cancer-vaccine-biontech.html.

[19] K. P. Venkatesh, K. T. Kadakia, S. Gilbert, Learnings from the first ai-enabled skin cancer device for primary care authorized by fda, npj Digital Medicine 7 (2024) 156. URL: https://www.nature.com/articles/s41746-024-01161-1. doi:10.1038/s41746-024-01161-1.

[20] T. I. S. I. Collaboration, Isic challenge, 2018. URL: https://challenge.isic-archive.com/data/.

[21] R. J. Friedman, D. S. Rigel, A. W. Kopf, Early detection of malignant melanoma: the role of physician examination and self-examination of the skin., CA: a cancer journal for clinicians 35 (1985) 130–151.

[22] B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm aversion: People erroneously avoid algorithms after seeing them err., Journal of Experimental Psychology: General 144 (2015) 114–126. doi:10.1037/xge0000033.

[23] J. M. Logg, J. A. Minson, D. A. Moore, Algorithm appreciation: People prefer algorithmic to human judgment, Organizational Behavior and Human Decision Processes 151 (2019) 90–103. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749597818303388. doi:10.1016/j.obhdp.2018.12.005.

[24] M. Schemmer, P. Hemmer, N. Kühl, C. Benz, G. Satzger, Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making (2022). URL: http://arxiv.org/abs/2204.06916.

[25] I. Ajzen, The theory of planned behavior, Organizational Behavior and Human Decision Processes 50 (1991) 179–211. doi:10.1016/0749-5978(91)90020-T.

[26] V. Venkatesh, H. Bala, Technology acceptance model 3 and a research agenda on interventions, Decision Sciences 39 (2008) 273–315. doi:https://doi.org/10.1111/j.1540-5915.2008.00192.x.

[27] B. F. M. for Economic Affairs, C. Action, Guidelines on the Protection of Health Data — bmwk.de, https://www.bmwk.de/Redaktion/EN/Dossier/guidelines-on-the-protection-of-health-data.html, 2025. [Accessed 29-03-2025].

[28] Q. Ha, B. Liu, F. Liu, Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge, arXiv preprint arXiv:2010.05351 (2020). URL: https://arxiv.org/abs/2010.05351v1.

[29] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: International conference on machine learning, PMLR, 2021, pp. 10096–10106.

[30] N. Papernot, P. D. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, CoRR abs/1803.04765 (2018). URL: http://arxiv.org/abs/1803.04765.

[31] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014. URL: http://arxiv.org/abs/1312.6034.

[32] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.