Concept-Driven Explainability Methods for AI in Medical Diagnostics

Bc. Jakub Šimko
Supervised by: Ing.Martin Dubovský

Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia

Abstract

Medical diagnostics, particularly in fields like histopathology, rely on the expert interpretation of complex visual data. As artificial intelligence (AI) systems become increasingly integrated into these processes, their lack of transparency presents a significant barrier to adoption. To bridge this gap, explainability methods are essential to ensure that AI-generated insights align with the mental models of medical professionals.

This study extends the End-User-Centered Explainable AI (EUCA) framework by introducing novel concept-based explainability methods and the Concept-Driven Design (CDD) methodology. The concept-based extension includes concept alignment and concept importance, which enhance the interpretability of AI predictions by ensuring their alignment with domain-specific concepts used by medical professionals. Furthermore, the CDD integrates human-centered design principles with explainability techniques to develop AI systems that better align with expert reasoning.

An initial round of expert testing demonstrated the usefulness of the proposed concept-based explainability methods. The preliminary results suggest a positive impact, highlighting the potential of these methods to improve AI-assisted medical diagnostics.

Keywords: Artificial Intelligence, Explainable Artificial Intelligence, Explainability Methods, Concept-based Explainability

1 Introduction

Advancements in machine learning (ML) promise the creation of Artificial Intelligence (AI) capable of perceiving, learning, deciding, and acting independently. However, these systems may not be able to explain their decisions and actions to human users. Therefore, Explainable Artificial Intelligence (XAI) is essential for achieving understanding, trust, and effective management of artificially intelligent assistants [3]. XAI is not just about an AI's ability to solve a problem with a certain accuracy, but also its capacity to explain why a particular solution was chosen. This research area aims to make the results of AI systems more understandable to humans [12]. However, these methods often fail to align with the way medical professionals reason about diagnoses, making AI explanations difficult to interpret in clinical contexts. The End-User-Centered Explainable AI (EUCA) framework attempts to bridge this gap by systematically categorizing explainability techniques and structuring them in a way that is more accessible to non-technical users. Although EUCA provides a strong foundation, there remains a space for further refinement to ensure that AI-generated explanations align with domain-specific high-level concepts used by medical experts.

This study extends the EUCA framework by introducing novel concept-based explainability methods tailored to the Nottingham Grading System (NGS) for breast cancer classification. We propose new methods that allow AI-generated explanations to be structured around clinically relevant concepts. This ensures that medical professionals can interpret and validate AI decisions more effectively. Furthermore, we propose a Concept-Driven Design (CDD) approach, which integrates human-centered design principles with explainability techniques to develop AI systems that align more closely with expert reasoning.

To evaluate the usefulness of these concept-based explainability methods, we conducted a testing with a domain expert. Our findings indicate that that concept-based explainability methods offer more meaningful insights than traditional explainability techniques, such as example-based or feature-based approaches. The findings highlight the usefulness of concept-based explanations in making AI-generated insights more interpretable and clinically relevant.

This paper is structured as follows: Section 2 discusses related work on explainability methods in medical AI. Section 3 details the research methodology. Section 4 describes the feature extraction process essential for developing the explainability methods. Section 6 introduces the proposed concept-based explainability methods, while Section 7 presents the expert evaluation results of the implemented explainability methods. Finally, Section 8 concludes with a discussion of the findings and directions for future research.

2 Related work

Several studies have explored different explainability techniques in medical AI applications. Calisto et al. [1] introduced the BreastScreening web application, which integrates an AI assistant to provide radiologists with a second opinion in breast cancer diagnosis. This solution employs a DenseNet model for classification and segmentation, allowing radiologists to accept or reject AI-generated classifications. The system includes an explanation functionality via heatmaps that visually indicate lesion severity. Testing with 45 clinicians demonstrated that diagnostic accuracy increased with AI assistance, and diagnoses were completed 31% faster. Furthermore, 93% of clinicians reported trust in the system's recommendations, highlighting the importance of integrating explainability into AI-driven medical tools.

Similarly, Tosun et al. [18] developed HistoMapr-Breast, an application leveraging XAI to enhance breast biopsy diagnostics through automated whole slide image analysis. A key contribution of HistoMapr-Breast is its ability to quantify diagnostic features and translate them into pathologist-friendly descriptions. The model identifies 18 different features which provides a pathologists with a clearer understanding of AI-driven decisions. The system includes a "Why?" button that allows pathologists to explore the rationale behind AI-generated classifications, providing explanations in an interpretable and clinically relevant manner. This feature highlights which attributes contributed to the prediction, their respective weights, and the model's confidence score. Additionally, the system compares the analyzed region with similar cases from a reference database, offering a broader context for decision-making.

Although previously mentioned studies have employed explainability methods that were available at the time, End-User-Centered Explainable AI (EUCA) synthesized a comprehensive taxonomy of explainability techniques based on an extensive review of prior research. Due to its structured and user-centered approach, EUCA was selected as the foundation for this study. EUCA specifically addresses the challenges of making AI understandable for non-technical users. One of the primary issues is the lack of AI-related knowledge among end-users, making traditional explanation techniques ineffective. Additionally, different user roles, tasks, and goals require tailored explainability approaches. To tackle these challenges, EUCA provides a structured framework comprising twelve user-friendly explanation forms, categorized into four main types. The **feature-based explanations** group includes feature attribute, feature shape, and feature interaction methods, which highlight key features influencing AI predictions. The **example-based explanations** category consists of similar, typical, and counterfactual examples, offering context-driven justifications for AI outputs. The **rule-based explanations** set includes rule text and decision tree visualizations, providing structured reasoning paths. Finally, the **contextual information** category offers insights into input data, output predictions, model performance, and dataset metadata. The EUCA framework also introduces a prototyping workflow that facilitates the development of explainable AI solutions.

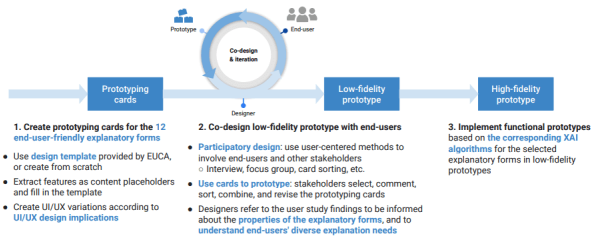

Figure 1: The suggested prototyping workflow using EUCA framework [6]

3 Research methodology

We chose to use the EUCA and its proposed workflow, introduced in section 2, for the prototyping of explainability methods, as it provides templates for 12 established explainability techniques and serves as a solid foundation for method design. Additionally, we extend the framework by incorporating new concept-based explainability methods that emerged through a deeper understanding of the domain.

The first step in the process for designing the explainability methods involved identifying the source of domain knowledge. Given the high value of domain experts' time and their significant time constraints, we chose to extract domain knowledge from existing scientific publications. By studying an extensive range of works focused on the Nottingham Grading System (NGS), referenced in section 4, we were able to acquire sufficient knowledge to initiate the first iteration of the EUCA card prototyping workflow, as depicted in fig. 1. Using the extracted domain knowledge, we identified specific features relevant to each of the NGS criteria. These features were then used to design low-fidelity and high-fidelity prototypes containing a total of 36 EUCA explainability cards, with 12 explainability methods dedicated to each criterion within the NGS. For example-based explainability methods, the feature extraction process focused on identifying the most relevant images from the NGS domain that could be applied to specific methods.

The proposed methodology for prototyping explainability methods, illustrated in fig. 2, not only guided our initial design efforts but also served as the foundation for extending the EUCA framework with novel concept-based explainability methods.

Figure 2: Our proposed methodology used for prototyping of explainability methods

4 Features extraction

Many ML studies and methods focused on classifying NGS criteria rely heavily on handcrafted features [4, 14, 5, 10, 8, 13]. These features are inherently interpretable for humans, which makes them particularly suitable for our prototypes of explainability methods.

For the **mitotic count** criterion, we concluded that interpreting why AI predicts a specific number of mitoses within a region is not particularly meaningful. Instead, the focus should shift to individual mitoses, providing explanations for why AI identified specific cells as mitotic. For this criterion, we identified more than seventy handcrafted features from the studies [14, 5, 10, 8]. These features can be broadly categorized into morphological features (e.g. area, eccentricity), texture features (e.g. chromatin density, contrast) and color and intensity features (e.g. mean intensity, skewness of patch intensities across seven color channels) [20]. However, to ensure that the selected features are easily understandable for domain experts, we chose three features that are both intuitive and seamlessly applicable to the EUCA explainability templates: eccentricity, aspect ratio and area.

The features of the **nuclear pleomorphism** criterion, as proposed by Faridi et al. in their study [4], include nuclear size, chromatin density, contour regularity, and the presence of nucleoli. Similarly, the study by Teoh et al. [17] generalizes the description of nuclear pleomorphism into three main categories: size, shape, and appearance. For our prototypes, we selected features from the size category, specifically nuclear size and nuclear aspect ratio, because of their ease of interpretation.

Tubule formation is defined as a structure consisting of at least one lumen (i.e., white region) surrounded by tumor cells [2]. To identify true tubules, Nguyen et al., in their study [13], utilized morphological features such as area, circularity, and curvature, as well as textural features like the histogram of intensity, histogram of gradient magnitude and orientation, and co-occurrence. Considering the need to interpret the extent of tubule formation within a specific area, we selected percentage of tubule formation and tubule aspect ratio as features for our prototypes of explanation methods.

5 Concept-based explainability

The concept-based explainability methods are built on Testing with Concept Activation Vectors (TCAV). TCAV is an interpretability method that leverages directional derivatives to quantify the influence of a user-defined concept on a classification result. For example, it can measure how sensitive a model's prediction of a zebra is to the presence of stripes [7]. In their study, Kim et al. [7] applied TCAV in the medical domain and observed that its results sometimes diverged from doctors' heuristics. This highlights TCAV's potential as a tool for assisting experts in interpreting and addressing model errors, particularly when predictions deviate from expert judgment.

5.1 Concept alignment

Concept alignment explanations help verify whether the model has learned to utilize abstract ideas that align with human reasoning. Rather than focusing on individual features, this method evaluates whether the model understands broader diagnostic concepts. The primary goal of this explainability method is to provide domain experts with a way to propose important concepts relevant to their field. Once a concept is proposed, the AI team employs TCAV to interpret the model's decisions.

The interpretation offers a distinct perspective for the domain expert. Instead of analyzing the explanation method solely from the model's perspective, trying to decipher what the AI communicates, the expert focuses on evaluating the alignment between their perception of the concept and the AI's perception and utilization of the same concept.

The high-fidelity prototypes for this explainability method are presented as follows: In the first image (fig. 4), we see an overview screen displaying all concepts suggested by the domain expert, along with their current status (submitted, in progress, or resolved). Domain experts can propose new concepts using the interface and once the proposal is submitted, the AI team begins working on the concept's interpretation. Upon completion, the domain expert can view the results, as displayed in fig. 5. On this screen, the domain expert evaluates the alignment between the proposed concept and the AI's interpretation.

Based on the testing conducted with the domain expert in section 7, this explainability method proved to be the most useful among all methods and served as the foundation for the development of the Concept Driven Design, described in section 6.

5.2 Concept importance

The concept importance explainability method demonstrates how a specific concept, aggregated from the concept-alignment method, influences the model's prediction. This explainability method, shown in fig. 3, helps

determine whether the model has learned to focus on relevant, high-level ideas or is being misled by unrelated concepts.

Concept importance is similar to the feature attribute explainability method from EUCA, which defines a feature as ”the individual pixels, a highlighted object, or an explicit concept”. The key difference lies in the source of the concept. In concept importance, the source is based on concept-alignment explanations, which rely on domain experts proposing the concept.

5.3 High-fidelity prototype

The concept-based explainability methods initially conceptualized in the low-fidelity prototype were further developed into a high-fidelity prototype and integrated into a modal interface. This interface is displayed to users when they choose to view explanations for a specific annotation. On the figures 3-5 is a curated selection of high-fidelity prototype screens showcasing our novel explainability methods.

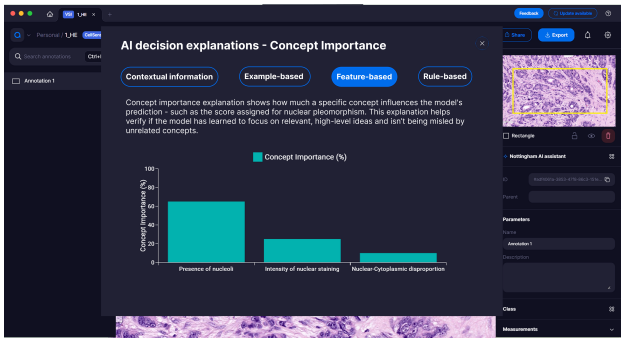

Figure 3: High-fidelity prototype of the screen containing the concept importance explainability method

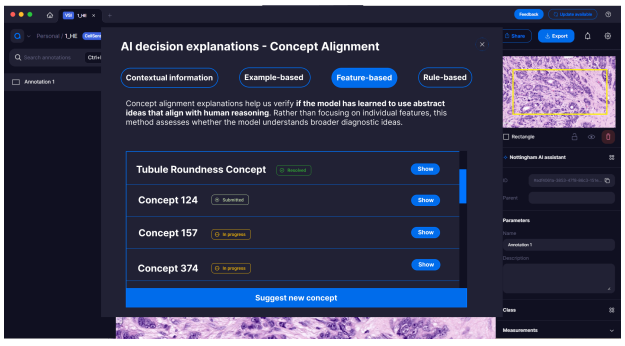

Figure 4: High-fidelity prototype of the overview screen of concept alignment explainability method

6 Concept Driven Design

Explanations that are overly technical or visually complex can be challenging to interpret, even for experts. While

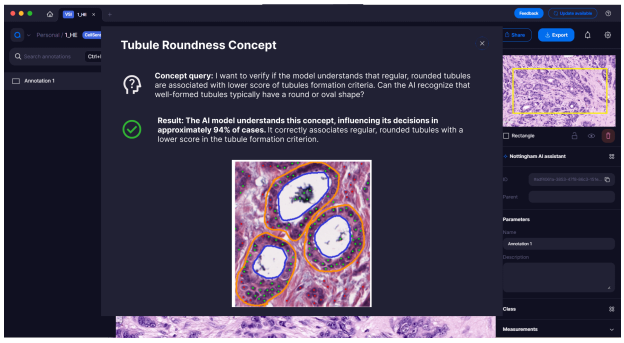

Figure 5: High-fidelity prototype of the screen with result of suggested concept for concept alignment explainability method

the intention behind most explanations is to facilitate user understanding, some methods are predominantly designed with a focus on technical execution rather than user experience. This technique centered design can impede users from fully grasping the explanations and may reduce their effectiveness [11]. When visual explanations are not intuitive or accessible, they can also undermine trust by creating confusion or misinterpretation [16].

The traditional approach to designing explainability, without utilizing the HCD approach, often relies on the limitations of currently available explainability techniques. While ML studies typically focus on extracting relevant input features from a model, along with their contribution to the output, which can serve as the foundation for explanations, they often fail to address whether these explanations are suitable for humans operating in the specific context [15].

The traditional ML approach also has one more critical issue. From the domain expert’s perspective, understanding the explanations provided by AI models can be challenging, as they must interpret these explanations solely from the AI’s point of view. In our case, this shift in perspective did not align with the expert’s mental model. Testing revealed that a more intuitive approach to explainability methods involves allowing the domain expert to interact with the AI by asking concept-related questions and assessing the alignment between their understanding of the concept and the AI’s interpretation. This interactive process empowers the expert to confirm whether the model employs the same concepts considered relevant in their domain.

We decided to expand this concept further by integrating it with the HCD approach, resulting in the first iteration of the CDD. Following extensive discussions with UX experts and multiple iterations, the final version of the CDD was developed, as illustrated in fig. 6. The proposed CDD, depicted in fig. 6, involves four key actors: the domain expert, UX expert, AI expert, and the AI model. It adheres to the HCD approach, with actions and objects color-coded to correspond to the specific phases of the HCD. The extension of the HCD lies in the inclusion of

a step focused on validation of the alignment between the domain expert's mental model and the AI's interpretation of concepts. For the questioning and mapping of mental concepts, resources such as the XAI question banks proposed by Liao et al. [9] or the conceptual framework for reason explanations introduced by Wang et al. [19] can be employed.

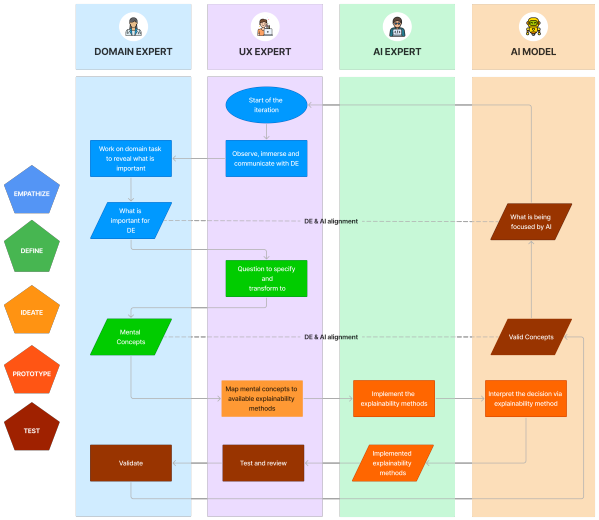

Figure 6: Flowchart illustrating the proposed CDD, integrating the HCD approach into explainability workflows. The framework involves four key actors: the domain expert, UX expert, AI expert, and the AI model, organized into distinct columns to highlight their individual contributions. The actions and objects within the flowchart are color-coded to correspond with the specific phases of the HCD shown on the left. The CDD emphasizes the validation of the alignment between the domain expert's mental model and the AI model's interpretation of concepts.

7 Testing with domain expert

The testing session with a domain expert in pathology employed the thinking aloud protocol and revolved around four key tasks. The primary aim was to assess the usefulness of the applied explainability methods and usability of the overall prototype design. Additionally, the testing aimed to gather valuable feedback to guide further iterations of the prototype or its eventual integration into the Annotaid application. A further objective was to observe the domain expert's interaction with the explainability methods, providing insights into potential areas for the development of new methods.

Task 1: Annotate a region using a square annotation tool. After completing this task, the expert was asked to evaluate the ease of annotation on a scale from 1 to 10, where 1 indicated a very easy process, and 10 represented the highest level of difficulty. **Task 2:** Use the AI assistant to determine the nuclear pleomorphism score. Upon finishing this task, the expert was asked to rate the ease of locating the AI assistant using the same 1 to 10 scale as in Task 1. **Task 3:** Review all explainability methods. In this task, the expert independently explored all the explainability methods provided in the prototype. Assistance was available if requested to clarify any uncertainties or ensure a complete understanding of the functionality of specific methods. This step was designed to prepare the expert for the subsequent evaluation. **Task 4:** Evaluate the usefulness of each explainability method on a scale from 1 to 10, where 1 represented not useful at all, and 10 represented highly useful.

The expert rated the first two tasks as very easy (score of 1 for both) and the usefulness ratings for the explainability methods, derived from tasks 3 and 4, are summarized in Table 1. Notably, the newly proposed **concept alignment** and **concept importance** methods received high ratings of 8, indicating that they significantly enhanced the expert's understanding of the AI model by aligning its reasoning with domain-specific diagnostic concepts. The **performance** method, which provides insight into the model's reliability via confidence scores, was rated moderately useful (5). In contrast, traditional methods such as saliency maps, example-based explanations, and feature-based techniques received low ratings, as the expert found these approaches less effective in understanding the AI's decision.

Explainability Method	Usefulness Rating (1-10)
Concept Alignment	8
Concept Importance	8
Performance	5
Decision Tree	5
Dataset	1
Similar Example	1
Typical Example	1
Counterfactual Example	1
Feature Attribute	1
Feature Shape	1
Feature Interaction	1
Rule Text	1

Table 1: Usefulness ratings for applied explainability methods

The expert commented that the concept alignment method was particularly valuable because it allowed her to verify that the AI model internalized diagnostic concepts in a manner that mirrored her own clinical reasoning. She noted, "It is much easier than trying to view the explanations solely through the AI's perspective." Similarly, the concept importance method was praised for providing localized insights into which features most influenced the predictions; however, the expert observed that its utility diminishes when applied globally. The performance method was appreciated for its role in calibrating trust in the AI

model, while the decision tree explanation was recognized for its simplicity in visualizing the decision process. In contrast, methods such as saliency maps and example-based explanations were found to be of limited practical value. Additionally, during the evaluation of the nuclear pleomorphism criterion, the expert suggested incorporating further concepts, such as the presence or absence of nucleoli and distinctions in the intensity of nuclear staining, to improve concept-based explanations. This instant feedback underscores the intuitive appeal and potential for rapid adoption of concept-driven methods in clinical practice.

8 Conclusion and Further Work

This study introduced novel concept-based explainability methods and explored the application of the extended End-User-Centered Explainable AI (EUCA) framework within the Nottingham Grading System (NGS) for breast cancer classification. By aligning AI-generated explanations with high-level, domain-specific concepts used by medical professionals, we extended the EUCA framework to bridge the gap between complex AI reasoning and clinical decision-making. Furthermore, we proposed a concept-driven design approach, integrating human-centered design principles with explainability techniques tailored to the needs of domain experts.

Expert evaluation revealed that concept-based explainability methods offer more meaningful insights than traditional explainability techniques, such as example-based or feature-based approaches. The findings highlight the usefulness of concept-based explanations in making AI-generated insights more interpretable and clinically relevant.

Despite these advancements, the practical implementation of concept-driven design in real-world clinical settings remains an open challenge. While this study establishes a theoretical foundation and presents a functional prototype, future research should focus on deploying and validating this approach. A key direction for future work is the iterative refinement of explainability methods through direct collaboration with domain experts, ensuring that AI-generated insights align with their evolving needs and expectations. Additionally, expanding the scope of concept-based explainability to other medical domains beyond breast cancer diagnostics could further validate its adaptability and effectiveness.

Ultimately, this study contributes to the broader effort of making AI-assisted medical diagnostics more interpretable and aligned with expert reasoning. By advancing concept-based explainability, we move closer to integrating AI systems that not only provide accurate predictions but also communicate their decision-making processes in a way that enhances usefulness of explainability methods utilized in AI assisted medical diagnostics.

References

[1] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. *Artificial Intelligence in Medicine*, 127:102285, 2022.

[2] Jean-Romain Dalle, Wee Kheng Leow, Daniel Racoceanu, Adina Eunice Tutac, and Thomas C Putti. Automatic breast cancer grading of histopathological images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3052–3055. IEEE, 2008.

[3] Gunning D Aha DW. Darpa's explainable artificial intelligence program. *AI Mag*, 40(2):44, 2019.

[4] Pegah Faridi, Habibollah Danyali, Mohammad Sadegh Helfroush, and Mojgan Akbarzadeh Jahromi. Cancerous nuclei detection and scoring in breast cancer histopathological images. *arXiv preprint arXiv:1612.01237*, 2016.

[5] Asmaa Ibrahim, Ayat Lashen, Michael Toss, Raluca Mihai, and Emad Rakha. Assessment of mitotic activity in breast cancer: revisited in the digital pathology era. *Journal of Clinical Pathology*, 75(6):365–372, 2022.

[6] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca: The end-user-centered explainable ai framework. *arXiv preprint arXiv:2102.02437*, 2021.

[7] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[8] Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical image analysis*, 45:121–133, 2018.

[9] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483*, 2021.

[10] Yao Lu, An-An Liu, and Yu-Ting Su. Chapter 6 - mitosis detection in biomedical images. In Mei Chen, editor, *Computer Vision for Microscopy Image Analysis*, Computer Vision and Pattern Recognition, pages 131–157. Academic Press, 2021.

[11] Shuai Ma. Towards human-centered design of explainable artificial intelligence (xai): A survey of

empirical studies. *arXiv preprint arXiv:2410.21183*, 2024.

[12] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[13] Kien Nguyen, Michael Barnes, Chukka Srinivas, and Christophe Chefd'Hotel. Automatic glandular and tubule region segmentation in histological grading of breast cancer. In *Medical Imaging 2015: Digital Pathology*, volume 9420, pages 92–98. SPIE, 2015.

[14] Monjoy Saha, Chandan Chakraborty, and Daniel Racoceanu. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Medical Imaging and Graphics*, 64:29–40, 2018.

[15] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, 2021.

[16] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 109–119, 2021.

[17] Chai Ling Teoh, Xiao Jian Tan, Khairul Shakir Ab Rahman, Ikmal Hisyam Bakrin, Kam Meng Goh, Joseph Jiun Wen Siet, and Wan Zuki Azman Wan Muhamad. A quantitative measurement method for nuclear-pleomorphism scoring in breast cancer. *Diagnostics*, 14(18):2045, 2024.

[18] Akif Tosun, Filippo Pullara, Michael Becich, D. Taylor, Jeffrey Fine, and S Chennubhotla. Explainable ai (xai) for anatomic pathology. *Advances in anatomic pathology*, 27:241–250, 07 2020.

[19] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[20] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003–034003, 2014.