EASL: A Framework for Designing, Implementing, and Evaluating ML Solutions in Clinical Healthcare Settings

Eric Prince Eric.Prince@CUAnschutz.edu
Computational Bioscience Program
Morgan Adams Foundation for Pediatric Brain Tumor Research Program
University of Colorado Anschutz Medical Campus
Aurora, Colorado, USA

Todd C. Hankinson Todd.Hankinson@ChildrensColorado.org
Department of Neurosurgery
Morgan Adams Foundation for Pediatric Brain Tumor Research Program
Children's Hospital Colorado
Aurora, Colorado, USA

Carsten Görg Carsten.Goerg@CUAnschutz.edu
Department of Biostatistics & Informatics
Morgan Adams Foundation for Pediatric Brain Tumor Research Program
Colorado School of Public Health
Aurora, Colorado, USA

Editor: Editor's name

Abstract

We introduce the Explainable Analytical Systems Lab (EASL) framework, an end-to-end solution designed to facilitate the development, implementation, and evaluation of clinical machine learning (ML) tools. EASL is highly versatile and applicable to a variety of contexts and includes resources for data management, ML model development, visualization and user interface development, service hosting, and usage analytics. To demonstrate its practical applications, we present the EASL framework in the context of a case study: designing and evaluating a deep learning classifier to predict diagnoses from medical imaging. The framework is composed of three modules, each with their own set of resources. The `Workbench` module stores data and develops initial ML models, the `Canvas` module contains a medical imaging viewer and web development framework, and the `Studio` module hosts the ML model and provides web analytics and support for conducting user studies. EASL encourages model developers to take a holistic view by integrating the model development, implementation, and evaluation into one framework, and thus ensures that models are both effective and reliable when used in a clinical setting. EASL contributes to our understanding of machine learning applied to healthcare by providing a comprehensive framework that makes it easier to develop and evaluate ML tools within a clinical setting.

1. Introduction

While progress has been made in the development of explainable artificial intelligence (XAI) methods for decision-support, their evaluation in a healthcare setting remains an open problem. AI is a general term for any computer system that can perform tasks that emulate

human intelligence (Moor, 2006). Machine learning (ML), which uses algorithms to learn from data and make predictions, decisions, and recommendations, is a key technology powering AI system (Tjoa and Guan, 2020). The field of XAI focuses on making AI systems more transparent and understandable (Markus et al., 2021). It utilizes techniques such as feature engineering, variable importance, and visualization to explain how an AI system is making decisions and predictions, thus making the system more interpretable for humans (Adadi and Berrada, 2018; Dwivedi et al., 2023).

It can be challenging to determine a clear definition for an explanation of an AI system (Verma et al., 2020). Different stakeholders may have divergent expectations of what constitutes a satisfactory explanation (Cai et al., 2019; Langer et al., 2021). Furthermore, the complexity of AI systems and the real-world contexts they are employed in can make it hard to convey their behavior in an efficient and comprehensive fashion. Explanations are often subjective and context dependent, making it challenging to generate a universal definition of what constitutes an explanation for an AI system. In general, there is a trade-off between the performance of AI models and their interpretability: the better they preform the less interpretable they are. When applied to high-risk settings, such as clinical healthcare, this trade-off can prevent theoretically powerful AI systems from being translated into real-world practice due to their lack of interpretability (Hatherley et al., 2022).

Explanations for AI systems often involve visualizations, metrics, and narrative descriptions (Tjoa and Guan, 2020). Visualizations such as bar graphs, heat maps, and decision trees can be used to illustrate the system's internal workings, while metrics like precision, recall, and accuracy can provide insight into its performance. Narrative descriptions can be used to explain the AI system's features, limitations, and potential applications (Gunning and Aha, 2019). Visual analytics (VA) methods, which combine computational approaches with interactive data visualization, offer an effective way to explore model outputs and communicate between an XAI system and the end-user. VA tools can integrate human judgment into algorithmic data-analysis by utilizing interactive visualizations (Cui, 2019). Recent advancements in VA methods tailored for the clinical setting are just beginning to emerge. A review of 71 VA platforms specifically developed for healthcare revealed that most platforms were created for classical and mainstream statistical methods (e.g., clustering and regression analysis), while few platforms were designed for predictive modeling (Ooge et al., 2022). Predictive modeling is a type of data analysis that applies data mining, machine learning, and statistical techniques to identify patterns and relationships in data, which can then be used to forecast or predict future outcomes or events. Such methods within an XAI system can be applied in healthcare contexts to gain better understanding of complex data and make informed decisions. However, a survey of 55 VA methods for clinical XAI highlighted the lack of examples of VA tools built for predictive modeling (Alicioglu and Sun, 2022). These works demonstrate that there is a need to develop methods for non-expert users to understand complex AI models.

Designing a clinical AI system requires a skilled multi-disciplinary team with expertise in various fields, such as clinical medicine, computer science, software engineering, artificial intelligence, design, observational studies, and other mixed methods, to overcome technical and humanistic challenges, such as tailoring a solution for the target audience and setting expectations for its function and benefits (Stephanidis et al., 2019; Quinn et al., 2021). Understanding the target audience, their goals, and the context of the decision-making

environment are all important considerations when determining if an explanation is understandable to a given person (Kann et al., 2021). Human-centered design (HCD) principles emphasize the importance of user involvement in the design process as well as iterative design and user testing to ensure that XAI solutions are tailored to the needs of users. However, there are several obstacles that prevent user involvement in healthcare (Chen et al., 2022a), such as (1) the knowledge gap between ML developers and the stakeholders in medicine, such as providers, administrators, or patients; (2) restrictions and ethical concerns limiting access to potential target users for iterative empirical tests in simulated setups for formative research or validation; (3) the complex nature of medical data (e.g., unstructured or high dimensional) and decision-making tasks from multiple data sources; and (4) the lack of ML developers' training in design thinking and human factors engineering. As a result, AI solutions are often incomprehensible to target users, which diminishes their relevance.

In this paper, we present the EASL (pronounced 'easel') framework to organize and execute the design, implementation, and evaluation steps necessary to translate AI to the clinical setting. EASL begins to address the four obstacles outlined above. We implemented EASL as a full-stack web-based multi-container Docker application and provide an example use case—a pilot study on the effects of XAI-assisted diagnoses on clinician decision-making. Through EASL, our goal is to support the ongoing efforts regarding XAI model development, and ML more broadly, by simplifying the clinical translation pipeline.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our contribution is not focused on the development or deployment of a specific machine learning model. Instead, the EASL framework tackles the overall process of designing, implementing, and evaluating machine learning models in clinical healthcare settings. As such, both the EASL framework itself and the lessons learned from applying EASL in the domain of image classification are generalizable to other models and applications in healthcare. By providing a unified environment for the design, implementation, and evaluation of XAI systems, our goal is to improve patient care by increasing physician access to information that can aid in their clinical decision-making.

2. Related Work

2.1. Human-Centered Design for Clinical XAI Development

Human-Centered Design (HCD) is becoming a topical solution for the development of clinical XAI tools. HCD for creating clinical solutions is a specific approach that puts patients and healthcare professionals at the center of the development and implementation process. It involves understanding users' needs and designing systems that are tailored to those needs, while also considering ethical, legal, and social implications. By using a human-centered design approach, clinical AI solutions can be designed to be more effective and to have a positive impact on patient care.

Schoonderwoerd et al. (2021) presented the DoReMi approach, a human-centered design workflow for AI-generated explanations in a clinical decision support system (CDSS) for diagnosing ADHD in children. DoReMi consists of domain analysis, requirements elicitation

and assessment, and multi-modal interaction design and evaluation. After analyzing the literature on the clinical diagnosis of ADHD, the group defined 20 information elements for XAI. Importantly, they considered the social contexts of the environment in which the CDSS would be used. These contexts include explaining the decision of the CDSS to a colleague that agrees or disagrees with the decision, explaining the decision to a parent that agrees or disagrees with the decision, and explaining whether the decision aligns with the user's own assessment or not. Using these elements and social contexts, they created explanation design templates and generated prototypes for user studies. Inspired by DoReMi, we designed EASL to be flexible and incorporate the approach into one centralized software system.

One of the most directly applicable resources for our example application of image classification is the INTRPRT framework, a set of HCD guidelines specific to medical imaging AI (Chen et al., 2022a). Like DoReMi, INTRPRT highlights the importance of formative user research, empirical user testing, general assessment of model transparency, and XAI systems for diverse stakeholders. The authors emphasize that these guidelines are a starting point and must be adapted and refined to the individual context of the AI application to ensure it is tailored to the end user's needs. The required refinement and adaption also assert that the successful implementation of XAI in the clinic merits customized HCD. By utilizing a centralized resource like EASL, AI developers, VA designers, and clinical end-users can all interact in one software environment, which simplifies our ability to co-create and adhere to HCD guidelines like DoReMi and INTRPRT.

2.2. Evaluating Clinical XAI

The evaluation of XAI is difficult in the clinical setting because of the complexities of clinical decision-making and the ambiguity of what constitutes a sufficient explanation to a given person. Doshi-Velez and Kim (2017) proposed a taxonomy for evaluating the interpretability of AI model explanations, which is especially applicable in the healthcare sector. The taxonomy includes application-grounded evaluation, human-grounded evaluation, and functionally-grounded evaluation (ordered by decreasing cost and complexity).

Application-grounded evaluation is the most rigorous framework as it evaluates a tool in the context of real users performing real tasks (Chen et al., 2022b); however, its implementation requires time, effort, and high standards of experimental design. Human-grounded evaluation is an attractive alternative for healthcare application, as it does not require access to the target community and can be conducted with lay people, providing a larger subject pool while also reducing cost. This type of evaluation is most useful for assessing more general aspects of explanation quality in healthcare settings. Functionally-grounded evaluations are another appealing option, as they do not require human experiments and instead use a formal definition of interpretability. These evaluations are especially beneficial when the class of models has already been validated via human-grounded experiments, or when human subject experiments are not possible or unethical. Additionally, this type of evaluation requires less time and financial resources than general human-subject experiments, and does not necessitate approval from an Institutional Review Board, making it a viable option for healthcare practitioners. Chen et al. (2022a) note that out of 68 clinical XAI manuscripts that they reviewed, only three performed an evaluation with end-users.

EASL is designed to handle the evaluation of all three levels of the Doshi-Velez and Kim (2017) taxonomy.

Chen et al. (2022c) proposed a use-case-grounded algorithmic evaluation called SimEvals to efficiently screen choices of information content and identify good candidates for a user study. Each SimEval involves training an agent to predict the use case label given the information content that would be presented to a user. The proposed method demonstrated that the agent's test set accuracy can be used as a measure of the predictiveness of information content (more specifically, of an explanation method) for each use case. The authors found that humans perform significantly better on explanations that SimEvals selects as promising compared to other explanations. We expect that this proposed work can be incorporated into future explanation evaluation workflows to enhance the efficiency and effectiveness of user study design.

In addition to academic research, there are many commercial examples of software for facilitating design and testing of visual analytics interfaces. These tools integrate with industry standard software like GitHub, Slack, and Google Analytics to facilitate version control, team communication, and real-time analytics of user studies. In addition, these tools provide functionality like email-based user surveys, in-app annotation, and tools for recruiting study subjects. However, these solutions are cost prohibitive for small scientific research groups and are not designed for handling medical data. EASL integrates web analytics functionality to provide an open-source platform to conduct comprehensive user studies without this additional cost.

3. Methods

In this section, we introduce the EASL framework, an end-to-end solution that facilitates the design, implementation, and evaluation of clinical AI tools. With its modular design, EASL is highly versatile and generalizable. It is applicable to a variety of clinical and biomedical contexts, from custom dashboards for independent scientists to complex AI systems deployed across multiple clinical sites. To illustrate how this framework can be applied in practice, we present EASL in the context of an example application: designing and evaluating a deep learning classifier to predict diagnoses from medical imaging.

3.1. Explainable Analytical Systems Lab (EASL)

As a conceptual framework, we modeled EASL after a potential artists' workflow by coining the design, implementation, and evaluation steps as the `Workbench`, `Canvas`, and `Studio` modules (Figure 1). The purpose of using the artist analogy is to emphasize the humanistic aspects and design focus when developing AI for healthcare. Each module contains design, implementation, and evaluation resources specific to a given project. A complete EASL project includes resources for data management, AI model development, VA development, service hosting, and usage analytics. Below we detail the purpose of each EASL module and provide specific examples regarding the functionality for each in the context of our example application.

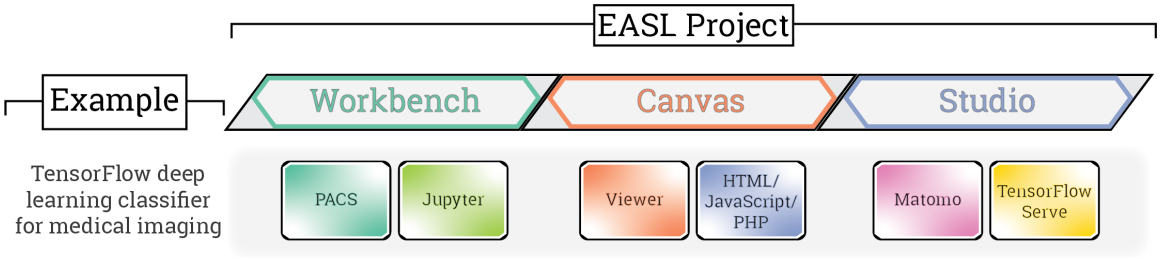

Figure 1: Graphical Overview of the EASL Framework. EASL projects are comprised of three modules (`Workbench`, `Canvas`, `Studio`). As an example, developing a deep learning classifier in TensorFlow for medical imaging requires a PACS (detailed below) and Jupyter (Python) environment in the `Workbench` module, medical imaging viewer and web development resources in the `Canvas` module, and web analytics (Matomo) and deep learning model hosting in the `Studio` module.

`Workbench`

The workbench is an important tool for artists as it allows them to prototype by exploring ideas, experiment with techniques, and bring their creative visions to life. Through the use of the workbench, artists can test out designs, colors, and forms to refine their projects. Similarly, prototyping for clinical AI tools involves creating a working model and experimenting with different features and functions to ensure the tool functions properly and accurately when used in a clinical setting. This process also helps identify potential problems, refine the user interface, and evaluate the performance and accuracy of the tool. In practical terms, developers typically manage data and develop initial AI models in order to identify and address any issues with the AI algorithm prior to its release. In the context of our example application, this module includes resources for handling medical imaging data and initial data evaluation and model development.

Medical imaging has a specialized form of data management due to the standardization of its data format, Digital Imaging and Communications in Medicine (DICOM). DICOM files are characterized by a distinct file format and a set of rules for exchanging information between medical imaging devices. These datasets are typically stored on a server, either in the cloud or on-premise, and accessed via a secure connection by authorized personnel. A Picture Archival and Communication System (PACS) is used to manage these large image datasets. PACS are medical imaging technologies that replace physical storage, such as film and x-rays, with digital images that can be viewed and shared over a network. PACS can be integrated with hospital information systems to give clinicians an overview of a patient's medical history. EASL provides a PACS for handling DICOM data and its associated software programs for accessing and managing the stored data within the database, powered by an Orthanc backend service (Jodogne, 2018).

After importing the medical images into a database, the next step of the process is to preprocess the data and begin developing AI models. This step is commonly performed by developers using an Integrated Development Environment (IDE). For our IDE, we utilize JupyterLab - an open-source web-based user interface for interactive computing. It pro-

vides a platform for data scientists, analysts, and developers to create and share documents containing live code, equations, visualizations, and narrative text. JupyterLab supports the use of multiple programming languages such as Python, R, Julia, and Scala. It also includes features such as a flexible file browser, an image viewer, and a multi-user server. Our implementation natively supports TensorFlow with NVIDIA GPU compute as well as standard Python packages like scikit-learn, pandas, matplotlib, and numpy. Developers can also install their preferred Python software packages such as PyTorch or plotly. Additionally, EASL provides a built-in API for easy accession of DICOM data from within the Python environments. Using this API, developers can import DICOM data into a format appropriate for AI development which greatly simplifies the logistics of this step.

`Canvas`

The canvas is the foundation for an artist's artwork. It provides a surface to work on and allows for a variety of textures, colors, and effects to be applied to create a unique piece of art. The canvas also serves as a physical representation of the artist's vision and can be used to express their individual style and creative vision. Prototyping for designing clinical AI tools is the process of creating a simplified version of the AI tool to test its functions and usability. This process allows developers to explore the design, refine the interface, and experiment with different features and functions prior to launching the full version. Through the iterative process for developing a prototype, developers can identify any issues that may arise before they become major problems, saving time and resources on development. In our example application, the canvas resources comprise a medical imaging viewer and a web development framework (e.g., HTML/JavaScript).

DICOM images can be viewed using specialized medical imaging software which allow users to view, manipulate, and analyze the images in order to make medical diagnoses or evaluations. The Open Health Imaging Foundation (OHIF) Viewer (Urban et al., 2017) supports viewing of standard medical image and enables developers to explore their databases with functionality like an integrated multi-view layout, contrast adjustment, measurements, and annotations. In addition, the OHIF Viewer component library (React Viewerbase) is available under an MIT copyright license. The MIT license is permissive and allows for modification, which enables the use of OHIF components in the visual analytics design process. The OHIF Viewer also has a rich library of extensions, which can readily be used to extend EASL to visualize whole slide microscopy data.

Web-based frameworks like PHP, HTML, and JavaScript are all highly flexible for designing visual analytics interfaces. PHP can create dynamic webpages and integrates with databases, HTML provides the visual layout and structure of the interface, and JavaScript supports coding interactive elements and creating custom data visualizations. Additionally, frameworks such as React, Vue, and Angular can be integrated to speed up the process of creating custom visual analytics interfaces. Libraries such as D3.js provide powerful data visualizations with minimal coding (Bostock, 2012). From a clinical implementation standpoint, a web-based framework for developing clinical AI has several advantages. First, it allows clinicians and researchers to collaborate more easily, since they can access and share data and results in real-time. Second, a web-based framework can provide a more secure environment for AI development, since the data and results are stored in a centralized lo-

cation and can be monitored for security breaches. Finally, a web-based framework can enable faster development and deployment of AI-based clinical solutions, since the data and results can be accessed from any device with an internet connection.

Studio

The purpose of a studio for an artist is to provide a space for them to create and showcase their artwork. A studio can be used to display finished pieces and facilitate interactions between the artist and their audience. It can also serve as a hub for networking and collaboration with other artists and art professionals. In the context of clinical XAI, the studio represents the functionality for hosting finished tools and evaluating their reception and use with the target audience. To implement this functionality we utilize web analytics which is the process of analyzing data related to a website or web-based application.

Hosting production AI models can be challenging due to their need for substantial computing resources, large datasets, and complex algorithms that can be difficult to manage. To address these challenges, hosting a binarized version of an AI model can be beneficial as it can improve inference speed and reduce latency, as well as reduce memory and storage requirements. Binarized models can be more energy efficient, as they require fewer computations than non-binarized versions, and can also improve model accuracy by pruning parameters and reducing overfitting. Furthermore, binarizing a model can increase security and reduce the risk of malicious actors exploiting vulnerabilities. To ensure accuracy, scalability, and reliability, AI models must also be constantly monitored and updated. Additionally, there are potential security risks associated with hosting production AI models, such as data breach and malicious actors, and AI models can often be expensive to develop and deploy, requiring substantial investments in hardware and software. In our example application we utilize TensorFlow Serve to host the AI model backend. TensorFlow Serve is a flexible, high-performance serving system for machine learning models, designed for production environments. It enables data scientists, developers, and production engineers to deploy new algorithms and experiments, while keeping the same server architecture and APIs.

Of the few examples of clinical AI tools that have been evaluated with clinical end users, they are predominately evaluated using survey-based methods. These methods have utility, but it is important to consider how the presentation of a survey could impact the study. For example, the use of a separate and often simplified software can lead to a scenario where the clinical user is not performing the task in a realistic environment. Therefore, we designed EASL to include Matomo (matomo.org, 2023), an open-source web analytics platform used to track and analyze visitor activity on websites. It offers features such as user segmentation and A/B testing. By integrating a small amount of code, developers of ML models can inject this web analytics functionality into any project and - importantly - can record user interaction directly within the user study session without needing additional software. In our example application, this resource enables direct time-stamped recordings of diagnostic predictions and time-stamped recordings of which DICOM slices were viewed for each prediction.

3.2. Implementation

EASL is a web-based application built using the Laravel PHP framework (Otwell, 2023), and deployed through a comprehensive Docker environment with middleware security. The EASL Base Image comprises a Linux operating system, MySQL backend, PHP support, and Matomo analytics integration. All the details pertaining to these modules and their implementation are outlined in the documentation. Laravel's Sail is used for managing services within the Docker development environment. With Docker, developers can package an application with all of its dependencies into a standardized unit, called a container. This allows applications to be quickly and reliably moved between development, test, and production environments, regardless of the underlying operating system or infrastructure. Additionally, Docker containers provide an isolated and secure environment for applications to run in, helping to reduce conflicts between applications running on the same system. Our implementation is publicly available on GitHub (`https://github.com/lericnet/easl`) under the GNU General Public License v3.0, providing developers with the flexibility to customize the framework and adding or removing Docker Images with minimal overhead.

4. Results

We applied EASL to replicate a clinical AI user study that we previously conducted without EASL. In this study, we designed and developed a deep learning classifier to predict diagnoses from medical imaging, and then evaluated the model in a healthcare setting with clinical experts. The replication of a previous study allows us to identify and discuss the effect that EASL had on the overall process of designing, implementing, and evaluating a ML solution in a healthcare setting.

Case Study: Impacts on Decision-Making Using Counterfactual Explanations of AI Diagnosis of Pediatric Brain Tumors from Neuroimaging

In this case study, we utilize EASL to test a hypothesis regarding how counterfactual AI visualizations can improve decision-making confidence while decreasing difficulty. We compare it to a previous iteration of the experiment in which we did not use EASL.

Clinical Background

Explaining AI-based predictions is fundamental for the development of clinical decision support systems. A common visual approach for explaining imaging data predictions is to overlay saliency maps onto images to allow users to interpret what visual features are associated with a given prediction. This approach can be difficult to utilize when differentiating nuanced concepts. For example, clinicians in neuro-oncology commonly must differentiate between a group of similar brain tumors (i.e., a radiographic differential diagnosis). We hypothesized that clinicians will be able to make diagnostic predictions more confidently, with less difficulty, and with greater accuracy if they are able to query "What are the most similar and dissimilar previously seen patients?" when given a novel case to diagnose. This concept is known as representativeness; a heuristic that clinicians use to interpret diagnostic data by considering similarity of a single case to a group of previously seen cases (Richie and Josephson, 2018).

Data and Tasks

We utilized a dataset of preoperative Magnetic Resonance (MR; n=52) and Computed Tomography (CT; n=61) image volumes and followed a human-centered design approach. We met with a board-certified neurosurgeon and neuroradiologist to conduct interviews regarding their workflow for diagnosing suprasellar tumors (Prince et al., 2020). We surveyed the literature to derive a library of visualization methods and discussed which techniques would be most relevant. Working with our collaborators, we defined three specific tasks for this interactive dashboard to perform:

1. *Standard neuroimaging interaction and interpretation.* Neuroradiologists utilize technology in many ways to acquire and interpret neuroimaging data. The goal of our design is to improve their existing workflow in a non-disruptive manner. Therefore, we must include standard functionality expected by practitioners like contrast/brightness adjustment, scroll-based z-axis navigation, multi-view synchronized views. The OHIF viewer provides a standard radiographic interpretation framework (Urban et al., 2017) that fulfills these requirements, see Figure 2.

Figure 2: Example of standard neuroimaging interaction interface (OHIF Viewer) shown with sagittal MRI test case and reference symbol at top right corner.

2. *Counterfactual representation of predictions.* We repurposed Google's What-If Tool (WIT; Figure 3) for counterfactual explanations (Wexler et al., 2019). Briefly, the WIT identifies counterfactuals using L_1 or L_2 norms in the output layer of a Ten-

sorFlow model. Typically, the WIT provides the functionality of modifying inputs; enabling the user to query "What if I change this feature value?" We disabled this functionality for our purposes to simplify the interface. Instead, users were only able to observe what similar and dissimilar previously seen patients were present in the dataset, thus providing the functionality of counterfactual matching. Additionally, there were a variety of other features present, which may have been overwhelming and confusing for unfamiliar users, as some of the terminology used was machine learning-based instead of healthcare-specific.

3. *Measurement of decision-making confidence, difficulty, and performance.* We adapted the ICE-T evaluation framework for our study to assess the visualization's value for our specific domain (Wall et al., 2018). Specifically, subjects would respond to "How confident are you in your prediction?" and "How difficult was your decision?" using a 5-point Likert scale for the prediction for each patient.

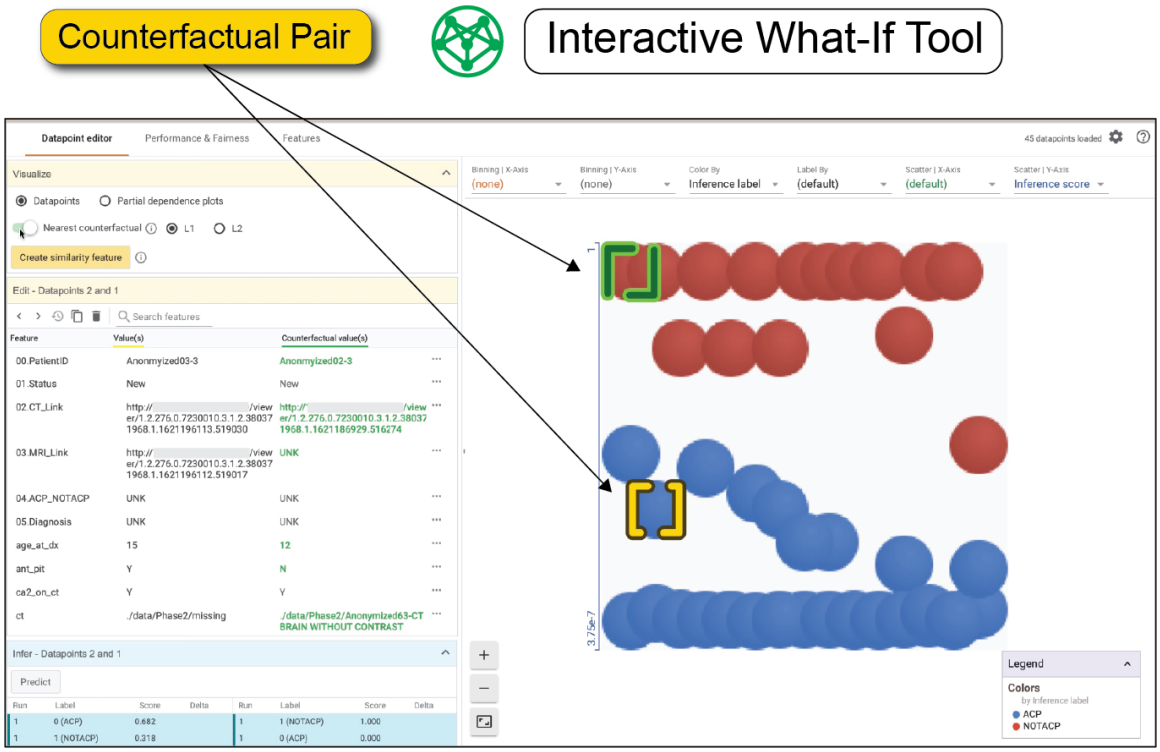

Figure 3: Example of WIT Interface that the participants utilized during the third phase of the study. The patient to diagnose is marked by yellow brackets, with the counterfactual match indicated by the green brackets. The points on the scatterplot are color-coded to show the predicted diagnosis group (ACP was blue and NOTACP was red). Metadata is located on the left side of the screen, with the predicted diagnosis scores listed in the bottom left table.

Technical Background

XAI methods are designed to explain an AI model's decision-making process. These explanations are often derived by perturbing features one at a time and monitoring change in some model performance metric, which then can translate to relative importance of that feature (Hailemariam et al., 2020). In the case of an image, each individual pixel (i.e., feature) is censored, classification is performed, and accuracy is reported. By comparing the change in accuracy for each pixel, we can determine a relative importance of that feature. Relative importance can then be overlayed as a heatmap on the original image, which allows the user to visually interpret what parts of the image contribute to the overall classification decision.

Different XAI methods accomplish this process using different mathematical approaches (Stepin et al., 2021). The most implemented XAI methods in biomedical research are LIME and SHAP (Hailemariam et al., 2020; Ribeiro et al., 2016; Štrumbelj and Kononenko, 2014). Continuing with the image example, LIME creates a surrogate model for a subset of images that an AI model considers to be similar and observes how changing features impacts the surrogate model. In simple terms, SHAP advances beyond that concept by using cooperative game theory to fairly allocate prediction importance across all input values. Importantly, these methods have been designed to explain these concepts to AI engineers and researchers, not to clinicians.

Study Design

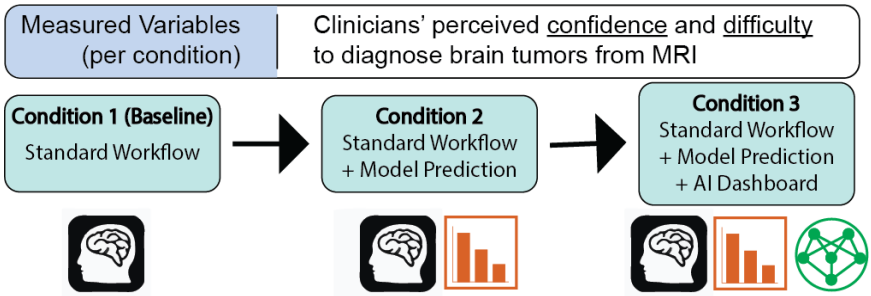

Figure 4: Schematic depicting each experimental condition and the variables measured. Symbols indicating which software components are included in each condition are depicted at the bottom.

We conducted a three-condition study with our two expert subjects, each subject performed all three conditions (Figure 4). For all study conditions, subjects were given a set (n = 28) of interactive PDF documents (Figure 5) that linked to the OHIF Viewer (Urban et al., 2017). Subjects were tasked with binary diagnostic prediction of adamantinomatous craniopharyngioma (ACP) versus other suprasellar tumors (NOTACP). In addition, subjects would respond to "How confident are you in your prediction?" and "How difficult was your decision?" using a 5-point Likert scale. For each patient, there was also a free response field for subjects to provide any additional feedback. The first condition provided only the OHIF Viewer. The second condition extends the first condition with a predicted

value. Predictions were generated using a deep learning model previously published (Prince et al., 2020). The third condition extends the second condition with the WIT. Subjects were prompted to think aloud during each condition and audio and screen recordings were captured for each session.

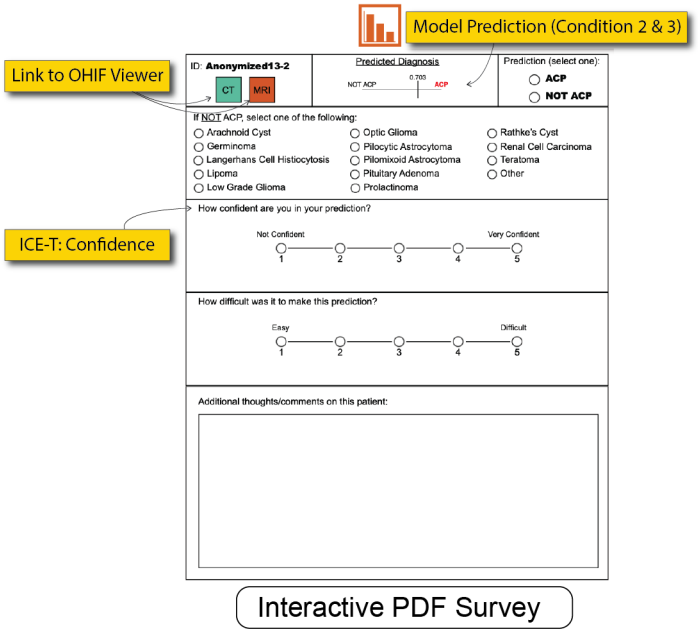

Figure 5: Example of interactive PDF with model prediction [document top center] clickable links to OHIF Viewer, radio buttons for reported prediction, survey questions, and free response text area.

Study Results

There was no effect on the diagnostic performance of clinical users (data not shown). There was no significant change in decision-making confidence and difficulty for each subject across the three study conditions for the NOTACP class of data (data not shown). However, there was a trend for increased diagnostic confidence and decreased diagnostic difficulty for both subjects with predictions for the ACP class of data (Figure 6). This trend was strongest for the third condition of the study.

Utility of using the EASL framework for conducting the study

We replicated our original experimental workflow using EASL. Data curation was streamlined to one access point using a standard File Browser interface, eliminating the need for unique data input requirements for the OHIF Viewer, model training, and WIT, which previously necessitated command-line interactions. In addition, the interaction between the DICOM server (which stores the images), the OHIF Viewer (which shows the images), and the WIT requires knowledge of network server protocols and relational databases. EASL

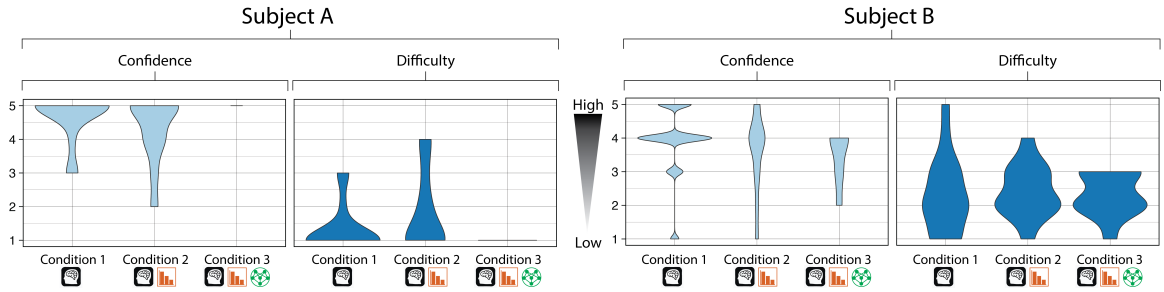

Figure 6: Study results: counterfactual XAI visualizations (condition 3) were associated with increased confidence and decreased perceived difficulty for diagnostic tasks. Each condition is shown on the x-axis with symbols indicating the respective components presents (standard interface, static model prediction, or WIT). The y-axis displays the 5-point Likert scale scores for the survey questions asked regarding decision-making confidence and perceived difficulty.

greatly simplified this interaction, allowing the researcher to focus on AI model and visualization development.

Utilizing our existing preferred workflows, there was no change in experience with respect to model development and testing within EASL, as this step was performed using the standard JupyterLab interface with TensorFlow software. The built-in API streamlined direct access to the DICOM data into a ready-to-use format. This process previously required many intermediary steps, and there was potential for human error. Similarly, there was no change in experience regarding the development of the visual analytics interface due to the use of the JupyterLab interface. Testing was greatly improved and simplified; previous iterations that required the use of a separate PDF building software and management of links on each document, were surpassed using simple HTML survey instruments that can be sent to and received from study participants via email.

5. Discussion

XAI has the potential to improve patient care by providing physicians with access to more comprehensive information to facilitate clinical decision-making. To ensure XAI solutions are tailored to user needs, the EASL framework makes it easier to implement HCD principles. User involvement and iterative testing throughout the development process are fundamental to HCD; however, user involvement and testing are also technically challenging and therefore often omitted, which leads to inadequate testing of proposed clinical XAI tools. EASL addresses this gap and encourages the development of XAI tools that meet the needs of clinicians, enhancing patient care through improved decision-making.

EASL streamlines the process of combining resources necessary for design, implementation, and evaluation of clinical XAI systems. Efforts related to the design, implementation, and evaluation of clinical XAI systems are often conducted separately, leading to siloed development. Additionally, there is an imbalance in the focus given to the

design and implementation stages compared to evaluation. To address these issues, we aim to formalize a comprehensive web development environment within the EASL framework that encompasses all three steps and places emphasis on evaluation.

The purpose of EASL as a framework is to establish a centralized space for the development of clinical ML solutions. Drawing inspiration from the commercial user experience design space, where software solutions facilitate real-time integration of design, development, and testing for web-based applications, we propose a similar approach for ML solutions in the clinical domain. EASL serves as a bridge to close the knowledge gaps between ML developers and healthcare stakeholders by providing a shared environment to store and utilize their knowledge. Furthermore, powered by web analytics, EASL facilitates iterative empirical testing and addresses restrictions and ethical concerns. It offers modular open-source solutions for medical data handling, implemented through Docker containers, and grants easy access to multiple data sources and deep learning functionality in Python.

In our example application, we conducted a pilot user study to examine how an interactive XAI tool can enhance a clinician's diagnostic performance. However, orchestrating this study proved complex due to the need to deploy separate local server instances for a PACS, a medical imaging viewer, and the What-If Tool with TensorFlow Serve instances of a deep learning model. The resources were coordinated through an interactive PDF created in Adobe InDesign. Clinicians commented that this testing environment was unrealistic and constrained their typical workflow. By consolidating these resources into a single computational environment, EASL simplifies the logistics of integrating the diverse resources required for designing, implementing, and evaluating clinical XAI systems.

The Dolshi-Velez and Kim taxonomy encompasses application-grounded evaluation, human-grounded evaluation, and functionally grounded evaluation. Both application-grounded evaluation and human-grounded evaluation involve human users and are supported by the complementary components of the complete EASL framework: the `Workbench`, the `Canvas`, and the `Studio`. These evaluations are typically conducted for relatively mature research prototypes. In contrast, functionally grounded evaluations focus on computational and/or algorithmic performance and do not involve human users. These types of evaluations are suitable for early-phase prototypes and can be facilitated by the EASL workbench and/or canvas.

The EASL web analytics resource provides a powerful tool for creating more realistic testing environments. By collecting comprehensive time-stamped results of how users interact with the system, we can create an accurate picture of the user experience. This type of data can support user studies on clinical XAI systems on a variety of endpoints, such as a laptop, desktop, tablet, or mobile phone. We can also measure whether the endpoint has an impact on the study results. Additionally, EASL collects specific information regarding the operating system, web browser, and screen size and resolution of each user's device. This information creates a more comprehensive picture of the user experience and ensures that the results of our studies are not skewed by the used endpoint. Taken together, this data allows us to create a realistic testing environment and to confidently draw conclusions about the effectiveness of our XAI systems.

EASL is well-suited for future clinical trials regarding XAI. Another future direction for EASL is to conduct a fixed endpoint study. For example, radiologists typically

interpret medical images on specialized DICOM-calibrated computer monitors which are optimized to communicate radiographic data through contrast and brightness channels. To ensure that potential decision-making improvement can be measured without constraining the environment, the endpoint for such a study should be the specialized equipment. Additionally, it is important to understand how different user endpoints impact decision-making when developing clinical XAI. As EASL is a web-based framework, it is flexible in terms of deployment and is thus suitable for future clinical trials regarding XAI.

Limitations The implementation of EASL presents various challenges that require expertise in full stack web development, artificial intelligence development, statistical programming, and hosting multi-container Docker applications. To address this and make EASL more accessible to non-expert programmers, we are developing project templates that include the necessary resources to quickly produce specific use cases. Additionally, we are working on simplifying the integration of computational resources needed for building clinical XAI tools and enhancing real-world translation. This involves focusing on the design, implementation, and evaluation aspects.

It is important to note that the current implementation of EASL is a web-hosting framework and not a publicly accessible resource. As a result, individual EASL instances do not communicate and can become highly customized, leading to interoperability issues. To overcome this limitation, hosting a persistent instance over a scalable cloud compute solution, such as Amazon Web Service, is a potential solution. However, this approach may lead to larger databases for XAI model development and could slow down small design studies with regulatory requirements.

In our Case Study, only two clinicians assessed the interface, which limits our ability to draw definitive conclusions from this application. However, we have ongoing work to expand the use case discussed in this manuscript to include a larger number of clinicians. EASL provides a framework that enables us to conduct these larger case studies effectively.

In terms of the generalizability of the EASL framework, the weakness lies in the fact that only one machine learning use case is tested, which is not sufficient to prove its generalizability. Therefore, we are currently undertaking additional studies to further demonstrate the framework's generalizability. In one such study, we leverage EASL for the development of custom single-cell RNA-sequencing analytics dashboards with built-in machine learning functionality. The intention is to follow this manuscript, which presents the EASL framework, with additional application papers across various biomedical domains.

6. Conclusion

We presented EASL, a powerful and versatile platform that can be used to design, implement, and evaluate XAI-based decision-support systems in healthcare. It can be utilized to create realistic testing environments, deploy resources, and collect comprehensive user analytics. EASL also allows for the scalability of the platform through the use of cloud computing solutions, facilitating further development in the field of XAI-based decision-support in healthcare.

Acknowledgments

This work was supported by NIH/NCATS Colorado CTSA Grant Number TL1 TR002533. Contents are the authors' sole responsibility and do not necessarily represent official NIH views. This work was also supported by the Morgan Adams Foundation for Pediatric Brain Tumor Research.

References

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Gulsum Alicioglu and Bo Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022.

Mike Bostock. D3.js - data-driven documents, 2012. URL http://d3js.org/.

Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW): 1–24, 2019.

Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1):156, 2022a.

Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. Use-case-grounded simulations for explanation evaluation. *arXiv preprint arXiv:2206.02256*, 2022b.

Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. Use-case-grounded simulations for explanation evaluation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022c. URL https://openreview.net/forum?id=48Js-sP8wnv.

Wenqiang Cui. Visual analytics: A comprehensive overview. *IEEE access*, 7:81555–81573, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

Yoseph Hailemariam, Abbas Yazdinejad, Reza M. Parizi, Gautam Srivastava, and Ali Dehghantanha. An Empirical Evaluation of AI Deep Explainable Tools. In *2020 IEEE Globecom Workshops (GC Wkshps*, pages 1–6, December 2020. doi: 10.1109/GCWkshps50303.2020.9367541.

Joshua Hatherley, Robert Sparrow, and Mark Howard. The virtues of interpretable medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics*, pages 1–10, 2022.

Sébastien Jodogne. The Orthanc Ecosystem for Medical Imaging. *Journal of Digital Imaging*, 31(3):341–352, June 2018. ISSN 1618-727X. doi: 10.1007/s10278-018-0082-y. URL https://doi.org/10.1007/s10278-018-0082-y.

Benjamin H Kann, Ahmed Hosny, and Hugo JWL Aerts. Artificial intelligence for clinical oncology. *Cancer Cell*, 39(7):916–927, 2021.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113: 103655, 2021.

matomo.org. Matomo analytics, 2023. URL http://matomo.org.

James Moor. The dartmouth college artificial intelligence conference: The next fifty years. *Ai Magazine*, 27(4):87–87, 2006.

Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. Explaining artificial intelligence with visual analytics in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1427, 2022.

Taylor Otwell. Laravel, 2023. URL https://laravel.com.

Eric W. Prince, Ros Whelan, David M. Mirsky, Nicholas Stence, Susan Staulcup, Paul Klimo, Richard C. E. Anderson, Toba N. Niazi, Gerald Grant, Mark Souweidane, James M. Johnston, Eric M. Jackson, David D. Limbrick, Amy Smith, Annie Drapeau, Joshua J. Chern, Lindsay Kilburn, Kevin Ginn, Robert Naftel, Roy Dudley, Elizabeth Tyler-Kabara, George Jallo, Michael H. Handler, Kenneth Jones, Andrew M. Donson, Nicholas K. Foreman, and Todd C. Hankinson. Robust deep learning classification of adamantinomatous craniopharyngioma from limited preoperative radiographic images. *Scientific Reports*, 10(1):16885, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-73278-8. URL http://www.nature.com/articles/s41598-020-73278-8.

Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. Trust and medical ai: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4):890–894, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, August 2016. URL http://arxiv.org/abs/1602.04938. arXiv: 1602.04938.

Megan Richie and S. Andrew Josephson. Quantifying Heuristic Bias: Anchoring, Availability, and Representativeness. *Teaching and Learning in Medicine*, 30(1):67–75, January 2018. ISSN 1040-1334, 1532-8015. doi: 10.1080/10401334.2017.1332631. URL https://www.tandfonline.com/doi/full/10.1080/10401334.2017.1332631.

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.

Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie YC Chen, Jianming Dong, Vincent G Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, et al. Seven hci grand challenges. *International Journal of Human–Computer Interaction*, 35(14):1229–1269, 2019.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Trinity Urban, Erik Ziegler, Rob Lewis, Chris Hafey, Cheryl Sadow, Annick D. Van den Abbeele, and Gordon J. Harris. LesionTracker: Extensible Open-Source Zero-Footprint Web Viewer for Cancer Imaging Research and Clinical Trials. *Cancer research*, 77(21): e119–e122, November 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0334. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5679226/.

Tejaswani Verma, Christoph Lingenfelder, and Dietrich Klakow. Defining explanation in an ai context. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 314–322, 2020.

Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. A Heuristic Approach to Value-Driven Evaluation of Visualizations. *IEEE transactions on visualization and computer graphics*, September 2018. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2865146.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2019.2934619. URL https://ieeexplore.ieee.org/document/8807255/.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665,

December 2014. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-013-0679-x. URL http://link.springer.com/10.1007/s10115-013-0679-x.