# Importance of User Control in Data-Centric Steering for Healthcare Experts

**Aditya Bhattacharya**[1], **Simone Stumpf**[2], **Katrien Verbert**[1]
[1]KU Leuven, Belgium
[2]University of Glasgow, Scotland, UK
aditya.bhattacharya@kuleuven.be, simone.stumpf@glasgow.ac.uk, katrien.verbert@kuleuven.be

## Abstract

As Artificial Intelligence (AI) becomes increasingly integrated into high-stakes domains like healthcare, effective collaboration between healthcare experts and AI systems is critical. Data-centric steering, which involves fine-tuning prediction models by improving training data quality, plays a key role in this process. However, little research has explored how varying levels of user control affect healthcare experts during data-centric steering. We address this gap by examining manual and automated steering approaches through a between-subjects, mixed-methods user study with 74 healthcare experts. Our findings show that manual steering, which grants direct control over training data, significantly improves model performance while maintaining trust and system understandability. Based on these findings, we propose design implications for a hybrid steering system that combines manual and automated approaches to increase user involvement during human-AI collaboration.

## Introduction

While Artificial Intelligence (AI) and Machine Learning (ML) have shown significant impact across various applications, high-performing prediction models alone are insufficient for effective human-AI collaboration (Rezwana and Maher 2022; Cai et al. 2019). Successful collaboration depends on factors beyond model performance, such as user understanding of the system's purpose and its advantages over existing methods (Cai et al. 2019). Limited user control over data instances and the lack of standardised guidelines for incorporating user feedback have further hindered human-AI collaboration (Wondimu, Buche, and Visser 2022). Although prior work has emphasised the need for better integration of end-user feedback (Chen et al. 2023), the role of user control in these interactions remains a debated topic (Zha et al. 2023; König 2022). Addressing these gaps is critical for advancing human-centred AI systems, especially for high-stake domains, such as healthcare.

To facilitate human-AI collaboration in healthcare, prior research highlights the benefits of involving healthcare experts in fine-tuning training datasets to improve prediction models (Bhattacharya et al. 2024b,a; Teso et al. 2022). We refer such an approach as *data-centric steering* as these approaches of finetuning prediction models with user feedback are also aligned with the principles of data-centric AI (Mazumder et al. 2023; Zha et al. 2023). Moreover, healthcare experts' domain knowledge is particularly valuable for identifying biases and limitations in training datasets, such as those found in patients' medical records, which could otherwise impact model performance (Bhattacharya et al. 2024a). Addressing these issues effectively has been shown to significantly improve prediction accuracy and fairness (Bhattacharya et al. 2024a; Feuerriegel, Dolata, and Schwabe 2020).

Despite its importance, little research has explored how varying levels of user control during data-centric steering influence model outcomes and collaboration with healthcare experts. This study addresses this gap by presenting two data-centric steering approaches: (1) manual steering and (2) automated steering, which healthcare experts can use for finetuning training datasets. Manual steering grants healthcare experts direct control over training data, enabling them to retain essential data points and predictor variables while removing problematic, corrupt, or irrelevant ones. In contrast, automated steering provides no direct control over the data but uses automated correction algorithms to identify and resolve potential data issues. In automated steering, experts can review these corrections and select methods with a single action, ensuring their consent.

We conducted a mixed-methods user study with 74 healthcare experts to examine the benefits and challenges of granting greater user control during data-centric steering through an interactive ML system. The process flow of the system is illustrated in Figure 1. Through the mixed-methods user study, we aimed to address two key research questions:

**RQ1.** What is the impact of involving healthcare experts in manual and automated steering on model improvement?

**RQ2.** How do manual and automated steering influence trust and system understandability for healthcare experts?

By focusing on these questions, we sought to explore the interplay between user control, model performance, and the collaborative dynamics of human-AI systems in healthcare. Considering the findings of this user study, our work

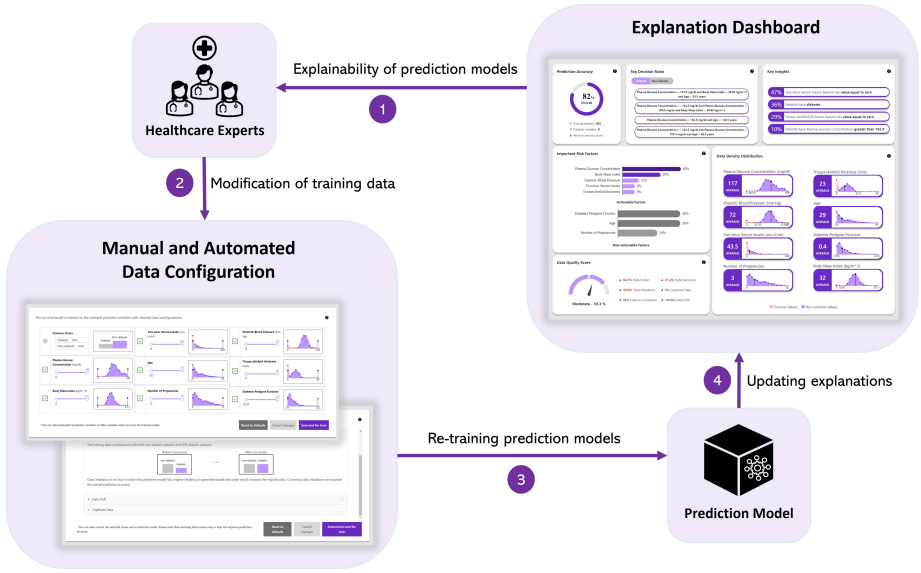


Figure 1: Process flow for our healthcare-focused data-centric steering system. (1) The system consists of a multifaceted explanation dashboard which combines data-centric and model-centric explanations for explaining the working of prediction models. (2) These explanations can facilitate healthcare experts to share their domain knowledge in modifying the training data for better model performance through manual and automated steering. (3) Prediction models can be re-trained using manual and automated steering on the configured training data. (4) After the prediction model is re-trained, the explanations are regenerated on the updated model and configured training data.

has three main contributions relevant to the field of human-centred AI:

1. **Empirical Contributions**: Our user study revealed that healthcare experts achieved higher prediction accuracy with manual steering, demonstrating its greater *effectiveness*. Despite the additional effort required during manual steering, it did not adversely affect user trust or system understandability. These findings highlight the potential benefits and trade-offs of empowering healthcare experts with greater control during steering, offering valuable insights into the design of such systems.
2. **Artifact Contributions**: We designed and developed a data-centric steering system that enables healthcare experts to apply their prior knowledge and configure the training dataset through manual or automated steering methods. The design, source code and technical architecture of our system are open-sourced on GitHub[1].
3. **Theoretical Contributions**: Our work presents a set of generic design implications for including healthcare experts in data-centric steering through manual and automated approaches. Additionally, these design implications can be extended to other application domains.

## Background and Related Work

### Data-Centric Steering

Historically, research in AI/ML has primarily focused on improving models rather than the quality of training datasets (Mazumder et al. 2023). While user-in-the-loop methods have been proposed to steer prediction models by adjusting algorithms, parameters and hyperparameters (Teso et al. 2022; Kulesza et al. 2015; Settles 2011; Cakmak and Thomaz 2011; Kulesza et al. 2010), these *model-centric approaches* face inherent limitations due to issues in the training data (Zha et al. 2023; Mazumder et al. 2023). Such steering approaches often lead to biased and inaccurate models mainly because of poor data quality (Bhattacharya et al. 2024a,b; Mehrabi et al. 2021). Recently, the growing recognition of the importance of data quality in AI systems has shifted attention toward *Data-Centric AI* (DCAI) (Zha et al. 2023; Mazumder et al. 2023), driving the development of methods to improve data quality to address issues due to biased, inaccurate and irrelevant predictions.

Feature selection (Li et al. 2017) is one such method utilised for preparing concise training data by selecting only important predictor variables instead of all the available variables. Data slicing is another method where a systematic approach is adopted to take a smaller segment of the data instead of the entire data (Zhang 2016). This method is particularly useful when the predictor variable has high outliers and skewed data distribution. Other methods include quality assessments (Sadiq et al. 2018) and quality improvements (Baylor et al. 2017). Quality assessments and improvement methods involve the detection and correction of common data issues like duplicate records, anomaly data points, correlated features, missing values, data drift detection, biased data, class imbalance and etc (Lones 2023; Ackerman et al. 2022; Kazerouni et al. 2020; Sadiq et al. 2018; Baylor et al. 2017). Despite differing sub-goals, these meth-

[1]https://github.com/adib0073/EXMOS


Figure 2: Screenshot of manual steering page that includes (1) feature selection control to include or exclude predictor variables and (2) feature filtering control to set the upper and lower limits for the predictor variables.

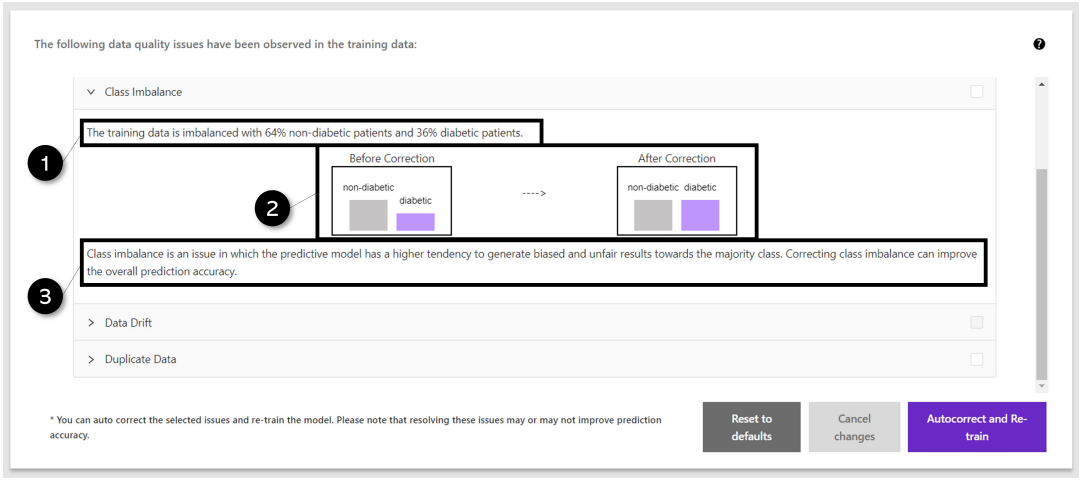


Figure 3: Screenshot of automated steering page that includes data issue explanations through (1) the quantified impact of these issues, (2) visualisations displaying before and after correction changes to the data or predictor variables, and (3) description of the issue and how its correction can impact the model performance.

ods share a common objective: improving training data quality for better prediction models (Zha et al. 2023).

To implement various DCAI methods for model steering, researchers have primarily relied on two approaches: (1) automation and (2) manual user collaboration (Zha et al. 2023). Automated algorithms are essential for handling the ever-growing volume of data in ML systems. For instance, data generation methods like Random Oversampling (Menardi and Torelli 2012), SMOTE (Chawla et al. 2002), and ADASYN (He et al. 2008) address class imbalance by generating synthetic samples for underrepresented categories. Similarly, automated algorithms can detect outliers, handle missing values, remove correlated features and duplicate records (Zha et al. 2023). Automation offer significant advantages, including reduced human error, improved efficiency, higher accuracy, and better reproducibility (Mazumder et al. 2023).

However, manual user involvement is critical in tasks where human expertise ensures the alignment of training data with domain-specific expectations (Zha et al. 2023). For example, manual efforts are vital for tasks like labeling (Zhang et al. 2022) and filtering data to remove noise or biases (Zha et al. 2023; Zhang 2016). Moreover, these manual and automated approaches are primarily designed for technical AI experts. Acknowledging the importance of involving domain experts with little to no AI knowledge in the AI solution pipeline (Bhattacharya et al. 2024a) and to examine the relative strengths of automated and manual approaches in improving training data quality, we conducted experiments comparing both methods with healthcare experts.

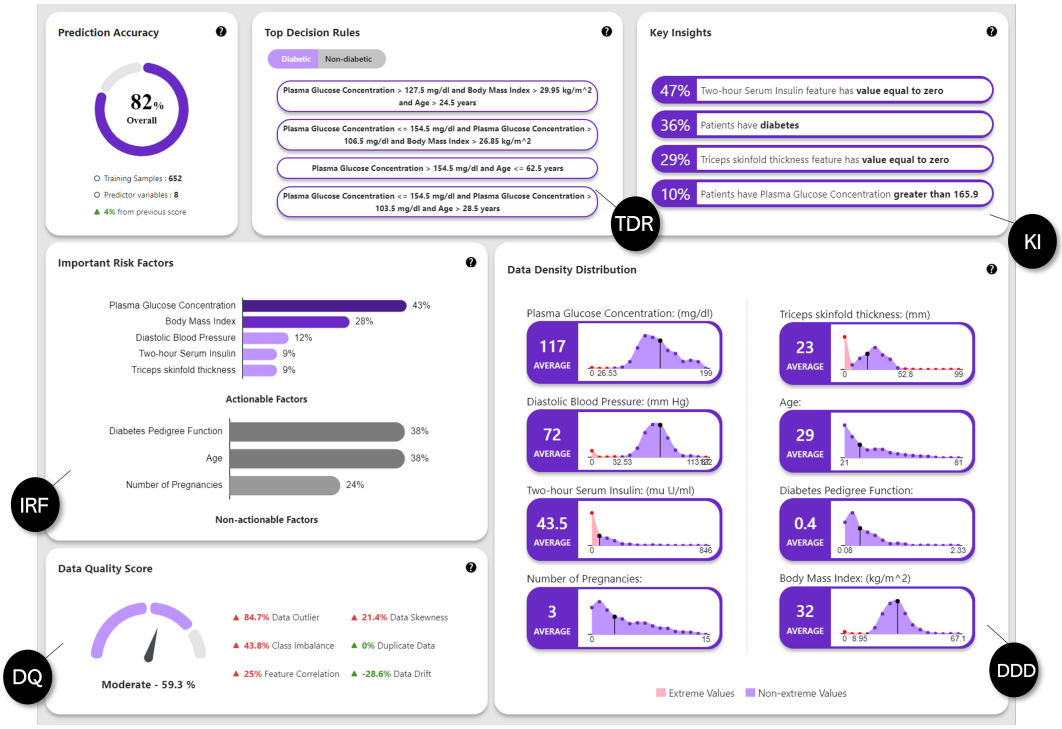


Figure 4: Screenshot of the explanation dashboard that includes the following visual components: Top Decision Rules (TDR), Key Insights (KI), Important Risk Factors (IRF), Data Quality (DQ) and Data Density Distribution (DDD).

## Explainable AI Methods

To effectively involve healthcare experts in the process of training, debugging and finetuning ML models, prior work has shown the importance of including Explainable AI (XAI) methods (Bhattacharya et al. 2024b, 2023; Lakkaraju et al. 2022; Adadi and Berrada 2018). Besides increasing the transparency of "black-box" algorithms, explanations can increase the understandability and trustworthiness of ML systems (Miller 2017; Liao et al. 2022; Bhattacharya et al. 2023).

Researchers have categorised explanation methods as *model-centric* or *data-centric*, based on the components of an ML system they address (Bhattacharya 2022). Model-centric approaches focus on evaluating the importance of parameters and hyperparameters within ML models, such as the SHAP-based feature importance method (Adadi and Berrada 2018; Lundberg and Lee 2017). In contrast, data-centric explanation methods analyse patterns in the training data to justify model predictions (Anik and Bunt 2021). These methods can summarise data patterns, detect biases and inconsistencies, and explain how issues like data drifts, adversarial attacks, or corrupted features affect model performance (Anik and Bunt 2021; Bhattacharya et al. 2023; Bhattacharya 2022).

While both approaches offer unique advantages and limitations, recent studies highlight that combining model-centric and data-centric explanations can provide more comprehensive and effective insights (Bhattacharya et al. 2024b; Demšar, Bosnic, and Kononenko 2019). In our steering approach, we incorporated both types of explanations in a dashboard to support healthcare experts.

## Steering System

***System Description***: Our data-centric steering system includes an explanation dashboard and a configuration page, enabling users to engage in either manual or automated steering (Figures 2 and 3 respectively), but not both simultaneously. It is designed specifically to support healthcare experts, such as doctors and nurses, in steering a diabetes prediction model for early detection of type 2 diabetes.

The design of the explanation dashboard (as shown in fig.4) was inspired by the work of Bhattacharya et al., who proposed integrating model-centric and data-centric explanations to improve user guidance in steering models. In line with their framework, the dashboard incorporates several visual components to improve user understanding. These include Top Decision Rules (TDR), which present decision rules generated by surrogate explainers; Key Insights (KI), which provide descriptive statistics from the training data; Important Risk Factors (IRF), which highlight feature importance using SHAP values; Data Quality (DQ), which summarises the estimated quality of the training dataset; and Data Density Distribution (DDD), which illustrates the frequency distributions of predictor variables to offer an overview of the dataset's characteristics. These elements collectively assist domain experts (like healthcare experts) in understanding and steering prediction systems.

***Prediction Model and Dataset*:** The diabetes predic-

tion model was developed using LightGBM algorithm (Ke et al. 2017), which was trained a type-2 diabetes detection dataset (Smith et al. 1988). This dataset comprises medical records of female patients and includes critical health information such as plasma glucose levels, body mass index, blood pressure, insulin levels, age, number of pregnancies, skinfold thickness, and pedigree function indicating the patient's family history of diabetes. A detailed data description and the various experiments conducted to train the prediction model are available in the supplementary material.

We selected this dataset for our experiments due to inherent data issues, including class imbalance (with significantly more non-diabetic than diabetic patients), numerous zero-values in predictor variables, and skewed data distributions. These characteristics made it ideal for investigating whether healthcare experts could identify and correct such issues to improve prediction models. Additionally, we avoided overly complex health datasets that could hinder system uderstandability and participant recruitment.

# Mixed-methods User Study

## Study Details

We conducted a mixed-methods between-subjects user study with 74 healthcare experts to compare manual and automated steering approaches. Participants were recruited from Prolific. On average, participants took 40 minutes to complete the study and were compensated at an hourly rate of $15 for their time. We have obtained the ethical approval for this study from KU Leuven with the approval number G-2021-4074.

The study included registered and in-training healthcare experts such as doctors, nurses, and paramedics, all of whom had prior experience in treating and caring for diabetic patients. Participants also self-reported familiarity with the predictor variables in the dataset and their relevance for predicting type 2 diabetes. To ensure balanced group assignment, participants were randomly divided into the manual and the automated steering groups, with 37 participants in each. Demographic information for each group is detailed in Table 1.

| | **Manual Group** | **Automated Group** |
|---|---|---|
| NO. OF PARTICIPANTS | 37 | 37 |
| AGE GROUPS | 20 – 29 years : 29<br>30 – 39 years: 4<br>40 – 49 years: 4 | 20 – 29 years: 31<br>30 – 39 years: 5<br>40 – 49 years: 1 |
| GENDER | Male : 19<br>Female : 18 | Male : 16<br>Female : 21 |
| EDUCATION LEVEL | Bachelor: 22<br>Master: 13<br>Doctorate: 2 | Bachelor: 28<br>Master: 7<br>Doctorate: 2 |
| HEALTHCARE EXPERIENCE | <1 year : 2<br>1 – 3 years : 13<br>3 – 5 years: 11<br>>5 years: 11 | <1 year : 2<br>1 – 3 years : 15<br>3 – 5 years: 10<br>>10 years: 10 |

Table 1: Participant information for the mixed-methods user study.

## Procedure

Participants were provided with detailed instructions outlining the study's objectives, their roles, responsibilities, and rights. After obtaining informed consent and collecting demographic information, we introduced the prototype using tutorial videos. These videos explained the usage scenario and provided an overview of the prediction model, explanation dashboard, and the steering mechanism assigned to each participant. The overall study flow is illustrated in Figure 5.

Following the tutorials, participants completed pre-task assessments, including objective understanding questions and perceived understandability and trust questionnaires. During this phase, they had full access to the explanation dashboard and their assigned steering method to assist in answering the questions. This step established a baseline for measuring how the steering process influenced their understanding and trust across the manual and automated groups.

Participants then completed a data-centric steering task using their assigned approach. Each participant had 10 minutes to modify the training data and retrain the prediction model, with the ability to configure the data multiple times to maximise model accuracy.

After the steering task, participants completed a post-task questionnaire to evaluate changes in their objective understanding, perceived understandability, and trust in the system. The post-task assessment also included open-ended questions to gather qualitative feedback on their experience. The system evaluation measures recorded during the study are elaborated in the next section.

## Evaluation Measures

To address our research questions, we collected the following evaluation measures in our user study. The complete study questionnaires are available in the supplementary material.

*Objective Understanding*: We evaluated participants' understanding of the system by measuring their objective mental model scores (objective understanding) before and after the steering task. This metric assessed participants' ability to identify key attributes driving the system's actions and predict outcomes based on changing conditions, following methods outlined by prior studies (Weld and Bansal 2018; Kulesza et al. 2015; Bhattacharya et al. 2024b; Cheng et al. 2019).

*Perceived Understandability*: In addition to objective understanding, we assessed participants' self-reported perceived understandability, or their confidence in predicting system behaviour, understanding its decision-making support, and using it effectively without detailed knowledge of its mechanisms (Hoffman et al. 2019). This was measured using a questionnaire from Hoffman et al., recorded both before and after the steering process.

*Perceived Trust*: Perceived trust, or participants' confidence in relying on the system, was evaluated using the trust scale from Jian, Bisantz, and Drury, recorded before and after the steering task to observe any changes.

*Post-Steering Model Accuracy*: We measured the updated prediction model accuracy after participants engaged in the

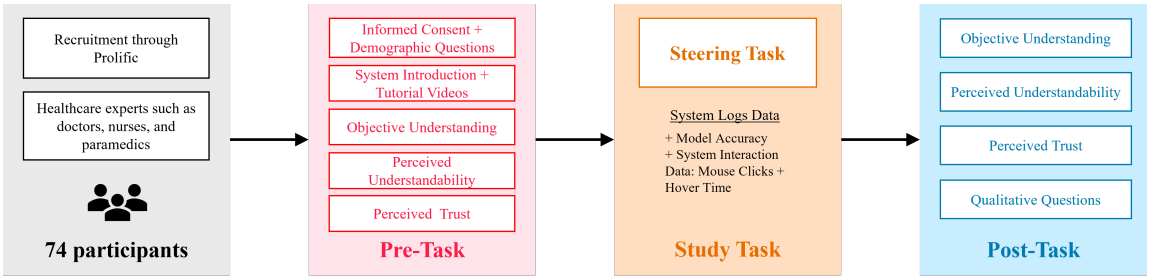


Figure 5: User study flow: this diagram illustrates the overall flow of our mixed-methods user study.

steering process, similar to Bhattacharya et al., to evaluate whether one group achieved better prediction accuracy improvements.

*Interaction Data*: System interaction data, such as mouse clicks, hover time, and model retraining attempts, were tracked to measure effectiveness and efficiency of manual and automated steering approaches, as per Verbert, Parra, and Brusilovsky. Effectiveness was calculated as the ratio of successful configurations (those that improved accuracy) to total configurations attempted, while efficiency was the ratio of total hover time to successful configurations.

*Qualitative Feedback*: Participants provided qualitative insights into their perceptions of understandability and trust in the system, which were used to identify qualitative factors influencing these perceptions.

### Data Analysis

As the quantitative data in our study did not meet normality assumptions, we employed the Mann-Whitney U-test to evaluate statistical significance between groups for the evaluation measures (McCrum-Gardner 2008). To assess changes in user understanding and trust before and after the steering task within a particular approach, we used Wilcoxon's signed rank test (McCrum-Gardner 2008). Additionally, Spearman's correlation coefficient (McCrum-Gardner 2008) was calculated to examine relationships between various measures, such as perceived trust, perceived understandability, objective understanding, and post-task prediction accuracy for each group. For qualitative data, we conducted thematic analysis following the approach proposed by Braun and Clarke to extract key themes from participant responses. To facilitate comparisons between manual and automated steering approaches, we used a range of plots and graphical visualisations to illustrate differences across the evaluation measures.

## Results

### What is the impact of involving healthcare experts in manual and automated steering on model improvement? (RQ1)

The Mann-Whitney U-test revealed that the manual steering group significantly outperformed the automated group in improving prediction accuracy ($U = 1054.0, p < .001$). Specifically, 84% of participants using manual steering achieved higher prediction accuracy than the default, compared to 67% of those using automated steering. From the system interaction data, manual steering required notably more effort in terms of click counts ($U = 1073.0, p < .001$). However, differences in average mouse hover time ($U = 706.0, p = .41$) and the number of model retraining attempts ($U = 734.0, p = .296$) were not statistically significant. In terms of *effectiveness*, the manual group achieved a significantly higher score (0.71) compared to the automated group (0.51, $U = 918.0, p = .005$). Conversely, for *efficiency*, the difference between the groups was not statistically significant ($U = 480.0, p = .27$). Boxed-violin plots and box-plots in Figure 6 illustrate the variations between the two groups across these measures, providing insight into RQ1.

### How do manual and automated steering influence trust and system understandability for healthcare experts? (RQ2)

The change in perceived trust between the manual and automated groups was not statistically significant, as indicated by the Mann-Whitney U-test ($U = 672.0, p = .896$). Both groups showed a slight increase in perceived trust after completing the steering task. However, the Wilcoxon signed rank test revealed no significant change in perceived trust from pre-task to post-task for either the manual ($W = 220.0, p = .79$) or automated ($W = 192.0, p = .40$) groups. Additionally, no significant correlation was found between post-task prediction accuracy and perceived trust for either group (manual: $r = 0.21, p = .20$; automated: $r = 0.05, p = .68$). Despite the manual group achieving higher model performance, their perceived trust levels were comparable to those of the automated group. Therefore, we conclude that the level of user control in model steering does not significantly affect perceived trust in the AI system.

The change in perceived understandability between the manual and automated groups was not statistically significant, as shown by the Mann-Whitney U-test ($U = 263.0, p = .09$). However, the automated steering group showed a greater average increase in perceived understandability (approximately 5%) compared to the manual group. Despite this, the increase in perceived understandability for the automated group was not statistically significant, as indicated by the Wilcoxon signed rank test ($W = 153.5, p = .103$). Similar to perceived trust and understandability, the Mann-Whitney U-test showed no significant difference in the change in objective understanding between the manual

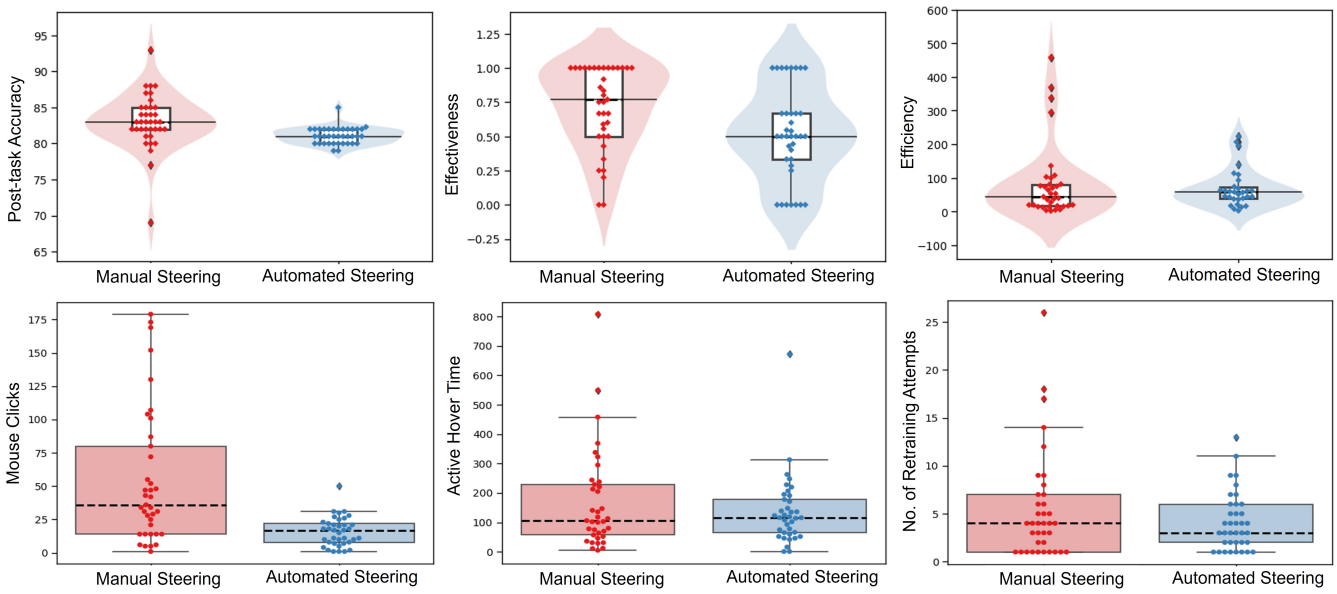


Figure 6: (Top row) Boxed-violin plot with marked data points showing the post-task accuracy scores, *effectiveness* and *efficiency* for manual and automated steering groups. (Bottom row) Box-plots showing variations in mouse-clicks, mouse hover time and number of retraining attempts for manual and automated steering groups.

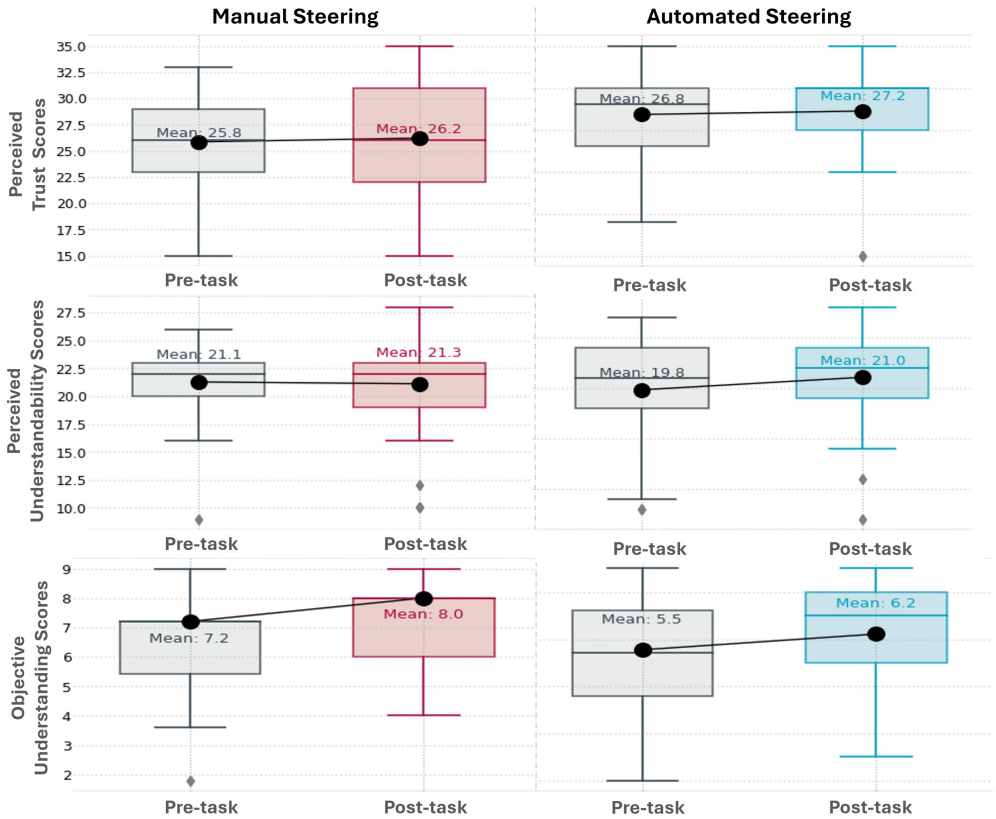


Figure 7: Comparing before and after scores of perceived trust, understandability, objective understanding for manual and automated steering groups.

and automated groups ($U = 646.5, p = .68$). However, the Wilcoxon signed rank test revealed a significant increase in objective understanding for the manual steering group ($W = 190.5, p = .019$). In contrast, the increase in objective understanding for the automated group was not statistically significant ($W = 263.0, p = .093$). Figure 7 presents box plots comparing the before and after scores of trust and understandability for both groups.

The results indicate that participants in both the manual and automated steering groups had similar levels of trust and understanding. To further explore the impact of control on trust and understandability, we analysed our qualitative data. Thematic analysis revealed key themes that offered deeper insights into how various factors in the steering process influenced user trust and understandability.

***Control is important, yet too much control is concerning***: Both groups valued the ability to configure the training data, which enhanced their trust in the model's results. While most manual steering participants gained trust through hands-on adjustments, some expressed concerns about maintaining data integrity. One participant noted: "*There is too much risk involved when dealing with medical data directly. Incorrect changes can be too risky.*" However, manual steering also helped participants better understand how changes to variables influenced predictions: "*After playing around with the configurations, I have a better understanding of the predictions.*"

***Explanations about data quality and data issues increase trust and reliability***: Both groups emphasised the importance of high-quality data for accurate predictions. Trust in the system grew when the explanation dashboard showed improved data quality after the steering process. As one participant stated: "*The data quality score gives me more confidence to trust the predictions. If it's lower, I'll be more sceptical and rely on my own judgment.*" Several also suggested that understanding the data collection process would further enhance trust: "*If the data quality is high, the system is more reliable, but knowing about the data source would increase my trust.*"

# Discussions

## Limitations

Before discussing the broader implications of our work, we acknowledge the following known limitations to ensure the transparency of our research:

1. *Potential limitations in the experimental prototype:* There are certain areas for potential improvement in our experimental prototype. Although the dataset used to train the prediction model fits all our experimental requirements, future research should consider investigating the impacts of different steering approaches on larger datasets. Additionally, integrating the explanation dashboard and the data configuration screens into the same view could enhance the system's usability and user interaction experience.
2. *Unexplored impact of manual and automated steering for other use cases*: The implications of this research work are limited only to classification models trained on structured datasets. Other use cases involving different types of datasets, such as image or text data, might require a different approach for manual and automated steering and were outside the scope of this research.

## Can We Depend Only on Manual Steering?

Our findings support providing healthcare experts with greater control through manual steering to achieve more *effective* model improvement. However, the manual approach comes with limitations. It increases the risk of human errors, particularly when users have only a partial understanding of the system. Such user-induced errors can degrade training data quality, which in turn affects model accuracy and diminishes user trust. Thus, our research raises an essential open question: "*What takes precedence for healthcare experts during data-centric steering: the pursuit of higher prediction accuracy or the assurance of better data quality?*" We argue that systematic collaboration between domain experts (like healthcare experts) and AI specialists is vital to balancing accuracy with data quality during manual steering. Such collaboration leverages domain knowledge alongside technical expertise, fostering informed decision-making that improves both model performance and the integrity of the data. We echo the recommendations from prior researchers (Bhattacharya et al. 2025; Zhang, Lee, and Carter 2022) who emphasised the importance of establishing a complementary partnership between domain experts and AI experts, especially in high-stakes domains, for building fair and responsible AI systems.

## Towards Hybrid Steering: Combining Manual and Automated Approaches

Manual and automated steering methods are not mutually exclusive in real-world applications. Given the distinct advantages observed in both approaches, we propose a hybrid system that integrates manual and automated steering to accommodate domain experts (like healthcare experts) who prefer varying levels of control. While our experimental setup did not incorporate hybrid steering as the primary objective of our study was to explore the comparative differences between manual and automated steering, future implementations should consider combining both approaches to maximise the benefits for healthcare experts.

## Involving a Group of Healthcare Experts During Manual Steering in a Practical Setting

Certain participants from the manual steering group expressed scepticism when their attempts to improve the model led to degraded performance instead of improvement. Such significant drops in prediction accuracy could adversely affect their trust in the system. To address this problem, we suggest replacing individual data-centric steering with a peer-approval and group consensus process similar to Bhattacharya et al. Leveraging the collective knowledge of a team of healthcare experts has the potential to overcome the limitations of individual feedback, such as human bias

and errors in training data (Mehrabi et al. 2021). By combining their expertise, a group of healthcare experts can achieve a more comprehensive understanding of the system and greater confidence in performing manual steering effectively. Moreover, involving groups of healthcare experts in steering sessions can help distribute cognitive load, making the process more manageable in time-constrained settings. In practice, we envision these sessions taking place collaboratively with AI experts, who can provide technical guidance and help mitigate any unintended effects on the prediction model resulting from domain expert interventions.

## Improving Inter-Stakeholder Collaboration in ML Systems Using Data-Centric Steering

ML systems often involve a variety of stakeholders with different backgrounds, such as developers, business leaders, policymakers, and users, who typically operate in isolation (Preece et al. 2018). While prior research has highlighted the importance of fostering collaboration across these groups (Preece et al. 2018; Nahar et al. 2022; Teso et al. 2022; Bhatt et al. 2020), the dynamics of collaboration between ML experts and domain specialists remain underexplored. Our findings suggest that both manual and automated steering approaches can play a key role in improving inter-stakeholder collaboration within ML systems. Building on Bhattacharya et al.'s guidelines, we propose a peer-approval process for steering systems that involves diverse stakeholders. This process leverages their collective technical expertise, domain knowledge, and practical experience to improve training data quality and develop more robust and contextually relevant ML models.

## Design Implications for Data-Centric Steering Systems For Domain Experts

To summarise the main findings and observations from our study with healthcare experts, we share the following implications for tailoring steering systems for domain experts:

- *Hybrid Steering for Balanced Control and Automation*: A hybrid approach that integrates manual and automated steering can help domain experts balance control and efficiency. While manual steering allows for domain expertise-driven adjustments, automated assistance can minimise human errors and improve workflow efficiency.
- *Facilitating Peer-Approval and Group Consensus*: Instead of relying solely on individual expert feedback, incorporating peer-review mechanisms among domain experts with diverse knowledge and background can improve the reliability of data-centric steering. This collaborative approach mitigates biases and errors, leading to more robust model adjustments.
- *Enhancing Explainability Through Interactive Visualisations*: Domain experts require clear explanations of model changes and their impact on predictions. Step-by-step interactive visualisations and explanations of training data quality that highlights the underlying issues can improve understanding and trust in the steering process.
- *Inter-Stakeholder Collaboration for Improved Data Governance*: AI systems for high-stakes domain (like healthcare) involve multiple stakeholders, including ML engineers, policymakers, and domain experts. Facilitating structured collaboration through shared decision-making frameworks can improve training data quality and model relevance.
- *Implementing Rollback and Version Control Mechanisms*: Inspired by version control systems like GitHub, steering systems should include rollback features that allow healthcare experts to track, revert, and audit changes. This ensures transparency and accountability in manual data modifications for domain experts.

Even though this work primarily focuses on understanding the perspective of healthcare experts, these implications are transferable to other application domains that require extensive domain knowledge for the development of responsible AI systems. We encourage future researchers working in human-centred AI to empirically evaluate these preliminary implications and refine them into concrete frameworks for robust data-centric steering systems for all domain experts with limited AI knowledge.

## Beyond Model Performance Improvement

While our study primarily focused on how data-centric steering impacts model performance, we believe its benefits extend well beyond accuracy metrics. Involving healthcare experts in the steering process can also support tasks such as bias detection and mitigation, as well as aligning training data more closely with real-world clinical scenarios. Prior work on AI bias and fairness (Bhattacharya et al. 2025; Mehrabi et al. 2021; Gianfrancesco et al. 2018; Feuerriegel, Dolata, and Schwabe 2020) highlights the importance of user involvement and domain expertise in addressing these broader concerns. Additionally, expert-guided data-centric steering can aid in the debugging and auditing of AI systems, helping ensure their compliance with ethical and societal standards established for responsible AI deployment (FRA 2019; Liao and Varshney 2022; Masís 2023; Szymanski, Verbert, and Vanden Abeele 2025). We envision that combining data-centric steering with explainable AI techniques can not only reduce model errors and bias, but also contribute to the development of more transparent, trustworthy, and fair AI systems.

## Future Work

Future research could investigate the combined effects of manual and automated steering. Additionally, conversational AI, such as chatbot-based interactions, could also improve system explainability and assist healthcare experts in creating more effective data configurations. To deepen understanding, future studies should engage diverse stakeholder communities to examine how manual and automated steering approaches influence inter-stakeholder collaboration. Prior research underscores the importance of rollback mechanisms, Kulesza et al. advocating for the reversible principle and Bhattacharya et al. highlighting the need to track user changes in interactive ML systems. Similar to version

control systems like GitHub, incorporating such features is strongly recommended to improve the usability and practicality of data-centric steering systems for domain experts.

## Conclusion

In conclusion, our research examined the potential for healthcare experts to improve prediction models using both manual and automated data-centric steering approaches. We developed a steering system for healthcare experts to guide a diabetes prediction model and conducted a mixed-methods user study with 74 participants. The study provided a detailed analysis of the benefits of each steering approach across multiple evaluation metrics. Our findings show that manual steering led to significantly higher prediction accuracy, demonstrating greater effectiveness. While manual steering required more effort, it did not substantially affect participants' trust or understanding of the system. Based on these results, we recommend granting more control to healthcare experts in fine-tuning prediction models. However, for practical applications, a hybrid steering approach that combines the strengths of both manual and automated methods would be ideal.

## Acknowledgements

We thank Ivania Donoso-Guzmán, Maxwell Szymanski, and Robin De Croon for providing helpful suggestions that improved this work. We also extend our gratitude to the members of the Faculty of Health Science, University of Maribor, Slovenia, for helping us with the exploratory study. This research was supported by the Flanders AI Research Program (FAIR) and Research Foundation–Flanders (FWO grants G0A4923N and G067721N.

## Positionality Statement

As researchers situated at the intersection of HCI and AI, we recognise that our perspectives are shaped by our technical backgrounds and institutional affiliations with access to advanced computational resources. While we involved healthcare professionals to ground our work in real-world practices, we are not clinicians ourselves. This outsider perspective influences our interpretations of needs and system usability for clinical practices. We acknowledge the importance of continued interdisciplinary collaboration to more accurately reflect the experiences of our target users.

## Ethical Considerations Statement

This study was conducted in accordance with the ethical guidelines of our institution and was approved through a multistage review process by the institutional ethics committee. The study design was carefully developed to uphold ethical standards from the outset. All participants provided informed consent prior to participation, and no identifiable data was collected; all responses were anonymised to ensure privacy. We took measures to minimise participant burden and emphasised their right to withdraw from the study at any time. In developing the steering system, we prioritised transparency and explainability to support responsible AI use in healthcare. Future work will continue to uphold these ethical principles, with a focus on inclusive and accountable system design.

## Adverse Impact Statement

While our work aims to improve human-AI collaboration in healthcare through data-centric steering, we acknowledge potential risks associated with these approaches. Granting domain experts direct control over training data may inadvertently introduce biases or lead to overfitting, particularly in the absence of rigorous validation by AI experts. Moreover, enabling user-driven steering could increase the system's vulnerability to adversarial manipulation or unintended misuse. To mitigate these risks, we recommend that such steering systems be deployed within controlled environments, with oversight from both AI and domain experts. Broad deployment should be preceded by thorough testing on representative datasets and ongoing evaluation to prevent the amplification of existing inequities or other unintended harms.

## References

Ackerman, S.; Farchi, E.; Raz, O.; Zalmanovici, M.; and Dube, P. 2022. Detection of data drift and outliers affecting machine learning model performance over time. arXiv:2012.09258.

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Anik, A. I.; and Bunt, A. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.

Baylor, D.; Koc, L.; Koo, C.; Lew, L.; Mewald, C.; Modi, A.; Polyzotis, N.; Ramesh, S.; Roy, S.; Whang, S.; Wicke, M.; Breck, E.; Wilkiewicz, J.; Zhang, X.; Zinkevich, M.; Cheng, H.-T.; Fiedel, N.; Foo, C.; Haque, Z.; and Jain, V. 2017. TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. 1387–1395.

Bhatt, U.; Andrus, M.; Weller, A.; and Xiang, A. 2020. Machine Learning Explainability for External Stakeholders. arXiv:2007.05408.

Bhattacharya, A. 2022. Applied Machine Learning Explainability Techniques. In *Applied Machine Learning Explainability Techniques*. Birmingham, UK: Packt Publishing. ISBN 978-1803246154.

Bhattacharya, A.; Ooge, J.; Stiglic, G.; and Verbert, K. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, 204–219. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701061.

Bhattacharya, A.; Stumpf, S.; Croon, R. D.; and Verbert, K. 2024a. Explanatory Debiasing: Involving Domain Experts in the Data Generation Process to Mitigate Representation Bias in AI Systems. arXiv:2501.01441.

Bhattacharya, A.; Stumpf, S.; De Croon, R.; and Verbert, K. 2025. Explanatory Debiasing: Involving Domain Experts in the Data Generation Process to Mitigate Representation Bias in AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; and Verbert, K. 2024b. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Braun, V.; and Clarke, V. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, APA Handbooks in Psychology®, 57–71. Washington, DC, US: American Psychological Association. ISBN 978-1-4338-1005-3.

Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Cakmak, M.; and Thomaz, A. L. 2011. Mixed-initiative active learning. *ICML 2011 Workshop on Combining Learning Strategies to Reduce Label Cost*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, V.; Bhatt, U.; Heidari, H.; Weller, A.; and Talwalkar, A. 2023. Perspectives on incorporating expert feedback into model updates. *Patterns*, 4(7): 100780.

Cheng, H.-F.; Wang, R.; Zhang, Z.; O'Connell, F.; Gray, T.; Harper, F. M.; and Zhu, H. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Demšar, J.; Bosnic, Z.; and Kononenko, I. 2019. Visualization of Explanations of Incremental Models. *Journal of Intelligent Computing*, 10: 121.

Feuerriegel, S.; Dolata, M.; and Schwabe, G. 2020. Fair AI. *Business & information systems engineering*, 62(4): 379–384.

FRA. 2019. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. In *FRA EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS*.

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11): 1544–1547.

He, H.; Bai, Y.; Garcia, E.; and Li, S. 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 1322 – 1328.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2019. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

Jian, J.-Y.; Bisantz, A.; and Drury, C. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4: 53–71.

Kazerouni, A.; Zhao, Q.; Xie, J.; Tata, S.; and Najork, M. 2020. Active Learning for Skewed Data Sets. arXiv:2005.11442.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 3149–3157. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

König, P. D. 2022. Challenges in enabling user control over algorithm-based services. *AI & Soc*.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. Atlanta Georgia USA: ACM. ISBN 978-1-4503-3306-1.

Kulesza, T.; Stumpf, S.; Burnett, M.; Wong, W.-K.; Riche, Y.; Moore, T.; Oberst, I.; Shinsel, A.; and McIntosh, K. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, 41–48. Leganes, Madrid, Spain: IEEE. ISBN 978-1-4244-7621-3.

Lakkaraju, H.; Slack, D.; Chen, Y.; Tan, C.; and Singh, S. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. arXiv:2202.01875.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2017. Feature Selection: A Data Perspective. *ACM Comput. Surv.*, 50(6).

Liao, Q. V.; and Varshney, K. R. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790.

Liao, Q. V.; Zhang, Y.; Luss, R.; Doshi-Velez, F.; and Dhurandhar, A. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. arXiv:2206.10847.

Lones, M. A. 2023. How to avoid machine learning pitfalls: a guide for academic researchers. arXiv:2108.02497.

Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874.

Masís, S. 2023. *Interpretable Machine Learning with Python: Build explainable, fair, and robust high-performance models with hands-on, real-world examples*. Packt Publishing, 2 edition. ISBN 9781803243627. Second Edition, 606 pages, Published October 31, 2023.

Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; Rojas, W. G.; Diamos, S.; Diamos, G.; He, L.; Parrish, A.; Kirk, H. R.; Quaye, J.; Rastogi, C.; Kiela, D.; Jurado, D.; Kanter, D.; Mosquera, R.; Ciro, J.; Aroyo, L.; Acun, B.; Chen, L.; Raje, M. S.; Bartolo, M.; Eyuboglu, S.; Ghorbani, A.; Goodman, E.; Inel, O.; Kane, T.; Kirkpatrick, C. R.; Kuo, T.-S.; Mueller, J.; Thrush, T.; Vanschoren, J.; Warren, M.; Williams, A.; Yeung, S.; Ardalani, N.; Paritosh, P.; Zhang, C.; Zou, J.; Wu, C.-J.; Coleman, C.; Ng, A.; Mattson, P.; and Reddi, V. J. 2023. DataPerf: Benchmarks for Data-Centric AI Development. arXiv:2207.10062.

McCrum-Gardner, E. 2008. Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, 46(1): 38–41.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Menardi, G.; and Torelli, N. 2012. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28: 92–122.

Miller, T. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences.

Nahar, N.; Zhou, S.; Lewis, G.; and Kästner, C. 2022. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, 413–425. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392211.

Preece, A.; Harborne, D.; Braines, D.; Tomsett, R.; and Chakraborty, S. 2018. Stakeholders in Explainable AI. arXiv:1810.00184.

Rezwana, J.; and Maher, M. L. 2022. Understanding User Perceptions, Collaborative Experience and User Engagement in Different Human-AI Interaction Designs for Co-Creative Systems. In *Proceedings of the 14th Conference on Creativity and Cognition*, C&C '22, 38–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393270.

Sadiq, S.; Dasu, T.; Dong, X. L.; Freire, J.; Ilyas, I. F.; Link, S.; Miller, M. J.; Naumann, F.; Zhou, X.; and Srivastava, D. 2018. Data Quality: The Role of Empiricism. *SIGMOD Rec.*, 46(4): 35–43.

Settles, B. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1467–1478.

Smith, J. W.; Everhart, J. E.; Dickson, W. C.; Knowler, W. C.; and Johannes, R. S. 1988. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

Szymanski, M.; Verbert, K.; and Vanden Abeele, V. 2025. *Human-Centered Explainable AI*. KU Leuven Doctoral Thesis: KU Leuven.

Teso, S.; Alkan, O.; Stammer, W.; and Daly, E. 2022. Leveraging Explanations in Interactive Machine Learning: An Overview. ArXiv:2207.14526 [cs].

Verbert, K.; Parra, D.; and Brusilovsky, P. 2016. Agents Vs. Users: Visual Recommendation of Research Talks with Multiple Dimension of Relevance. *ACM Trans. Interact. Intell. Syst.*, 6(2).

Weld, D. S.; and Bansal, G. 2018. The Challenge of Crafting Intelligible Intelligence. arXiv:1803.04263.

Wondimu, N. A.; Buche, C.; and Visser, U. 2022. Interactive Machine Learning: A State of the Art Review. arXiv:2207.06196.

Zha, D.; Bhat, Z. P.; Lai, K.-H.; Yang, F.; Jiang, Z.; Zhong, S.; and Hu, X. 2023. Data-centric Artificial Intelligence: A Survey. arXiv:2303.10158.

Zhang, J.; Hsieh, C.-Y.; Yu, Y.; Zhang, C.; and Ratner, A. 2022. A Survey on Programmatic Weak Supervision. arXiv:2202.05433.

Zhang, Q.; Lee, M. L.; and Carter, S. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Zhang, Z. 2016. Missing data imputation: Focusing on single imputation. *Annals of translational medicine*, 4: 9.