Advancing Medical Decision-Making through Human-AI Collaboration

Marc Goutier[1], Tizian C Dege[2], Maurice Rohr[2], Alexander Benlian[1], Christoph Hoog Antink[2]
[1]ISE - Information Systems and Electronic Services,
[2]KIS*MED - AI Systems in Medicine,
Technical University of Darmstadt, Darmstadt, Germany

Abstract

Integrating artificial intelligence (AI) into decision-making is most effective when AI complements human weaknesses and enhances strengths. Our study explores this by training a specialized AI model to classify ECG signals that are difficult for humans. We evaluated its performance against a baseline model and conducted an online study to assess how nudging strategies influence human reliance on AI suggestions. Results show that training AI on human-difficult cases boosts performance in resource-limited settings, maximizing its complementary potential. Nudging strategies also significantly impact collaboration, with effectiveness depending on task complexity and the accuracy of both AI and humans. Our Intelligent Nudging approach improved human-AI collaboration accuracy by 20%. Embedding complementarity into AI training and using tailored nudges presents a practical path to enhance decision-making, especially in critical domains like medical diagnostics. These findings offer valuable insights for designing effective human-AI teams that leverage each party's strengths for better outcomes.

1. Introduction

In recent years, artificial intelligence (AI) has demonstrated remarkable performance across a wide range of applications [1]. However, in clinical settings, where AI suggestions may have life-or-death consequences, reliance (i.e., if the human follows the suggestion of the AI or not) on AI must be approached with care. To improve human-AI collaboration, recent research has focused on explainable AI (XAI), which includes methods aimed at increasing the transparency of machine learning models [2]. While XAI improves our understanding of AI decision-making, it does not directly tackle the challenges of human decision making or the optimization of overall performance in human-AI collaboration. While humans and AI have their own capabilities in medicals tasks, the collaboration between humans and AI is far from optimal. As recent research found out, this is mainly because humans struggle to assess their capabilities compared to the AI and therefore fail to properly recognize when making a medical decision on their own or rely on the AIs suggestion [3]. In contrast, AIs are better than humans to assess the capabilities.

To address this, we propose *Intelligent Nudging* as a strategy to improve AI reliance in challenging decision-making contexts. Intelligent Nudging aims to subtly guide users toward more effective AI-collaboration, particularly in high-stakes environments. In this study, we explore the potential of Intelligent Nudging to improve human-AI collaboration in medical diagnostics. Using data from the PhysioNet Challenge 2017 [4], we categorized electrocardiograms (ECGs) into "easy" and "difficult" signals based on their classification difficulty for humans. To further study AI reliance, we conducted a two-phase survey with 96 participants who had medical training or ECG understanding. A pre-study with 27 participants validated our classification of difficulty levels. Afterwards, the main study compared an Intelligent Nudging group to a control group to assess how different forms of AI nudges affect the performance of the human-AI collaboration, measured by the accuracy of correctly classified ECG signals. In the main study, participants did not interact with an actual AI model; instead, we simulated AI assistance to explore how AI-generated suggestions could be presented effectively.

2. Methods

For the study design, the PhysioNet 2017 dataset was analyzed, and criteria for identifying easy and difficult samples were defined. From now on easy and difficult datasets describe the difficulty for humans to classify this data, unless otherwise stated. This section introduces our approach to detecting uncharacteristic data points within the dataset. Furthermore, nudging as a tool to steer human-AI collaboration is presented, followed by its implementation in an online survey to evaluate its effectiveness.

2.1. Dataset

For this study, we use the PhysioNet Challenge 2017 dataset, as we assume that single-lead ECGs are generally

straightforward for humans to classify [4]. Additionally, benchmark models are available, which we can use to evaluate AI-complementarity.

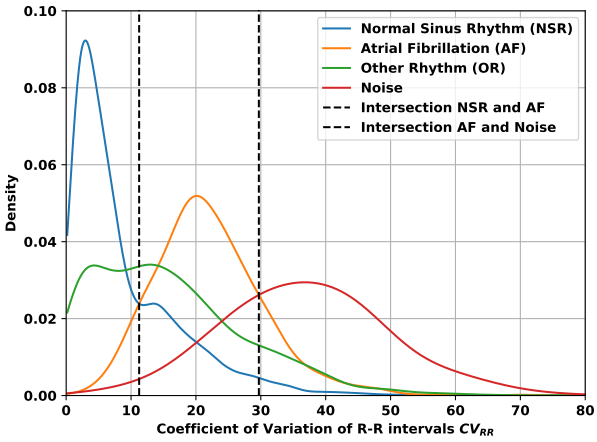

Figure 1. Kernel density estimation of coefficient of variation of R-R intervals for each class.

The dataset consists of four classes: Normal Sinus Rhythm (NSR), Atrial Fibrillation (AF), Other Rhythm (OR), and Noise. To distinguish between easy and difficult samples, we analyze the coefficient of variation of R-R intervals (CV_{RR}) using kernel density estimation. This allows us to identify peculiar data points that deviate from typical patterns.

Figure 3 shows the kernel density plot of CV_{RR} for all ECG signals in the dataset. The coefficient of variation CV_{RR} is calculated as the ratio of the standard deviation (SD_{RR}) to the mean R-R interval ($\overline{\mathrm{RR}}$):

$$CV_{\mathrm{RR}} = \frac{SD_{\mathrm{RR}}}{\overline{\mathrm{RR}}} \tag{1}$$

To differentiate between easy and difficult samples, we use the intersections between the density plots as boundary markers. These boundaries, highlighted in Figure 3, delineate distinct regions within the data distribution. Moreover, the plot exhibits three distinguishable peaks corresponding to three classes in the dataset. Data points that fall outside the defined boundaries are considered atypical and classified as difficult samples. *OR* shows no distinguishable peak, therefore we consider all *OR* signals as difficult data points. For the online survey, we only consider the classes NSR, AF, and OR. To confirm our assumption about easy and difficult data, we did an in-depth pre-study.

2.2. AI Complementarity

To test AI complementarity, we trained a model on two different datasets, both of which are subsets of the complete PhysioNet 2017 dataset. To maintain the overall distribution of the dataset, we applied stratified sampling. For the complementary model, data points with atypical R-R intervals were preferred, as described in Section 2.1, while randomly selected data points were used for training a baseline model. To ensure comparable results, all datasets contained the same number of samples. We performed an 80/10/10 split for training, validation, and testing, and evaluated the models on the test set. The model architecture used was ResNet1D by Hong et al., the winning entry of the PhysioNet Challenge 2017 [5]. We trained the models using the Adam optimizer and CrossEntropyLoss as the loss function. Both models were trained for 50 epochs, with early stopping employed as a countermeasure to overfitting. The results on the test sets are presented in Table 2.

2.3. Nudging

As explained, AI is better than humans at determining if the human or the AI should make the decision [3]. Therefore, we propose Intelligent Nudging as a means to leverage this strength of the AI while keeping the human in the loop. The concept of Nudging refers to influencing human behavior without restricting available choices [6]. While nudging can effectively guide human behavior toward better human-AI collaboration, its traditionally static nature limits adaptability in scenarios where tasks vary—shifting whether the human or AI is better suited to decide. To overcome this, we propose two key ideas for how the AI should nudge the human: (1) Nudging strategies should vary depending on the relative capabilities of the human and AI; (2) Nudging should be applied intelligently, i.e., task-dependently, to optimize overall accuracy.

In our main study, we implemented an AI able to give a suggestion to the human for a specific case, but the human always makes the final decision. We then vary the intensity of assistance based on relative accuracy. In cases where humans are predicted to outperform the AI, the nudge should be non-intrusive, so AI assistance is mostly ignored. When the AI is predicted to be better, it should use an intrusive nudge, so humans use the provided assistance with higher likelihood.

We used four different nudging types in our experiment. We vary the invocation of assistance, since invocation strongly influences whether a human follows it [7]. The four nudging types are, ranked by increasing intrusiveness:

- **Control**: The human does not see the assistance of the AI.
- **Reactive Nudge**: The user must explicitly request AI assistance to receive a suggestion.
- **Proactive Nudge**: The AI provides a suggestion by default without user solicitation.

- **Preselection Nudge**: The AI presents a suggestion and pre-fills the response form accordingly.

We define Intelligent Nudging as the adaptive application of these strategies based on varying tasks. Specifically, the AI should nudge the human more intrusively when its expected accuracy exceeds that of the human.

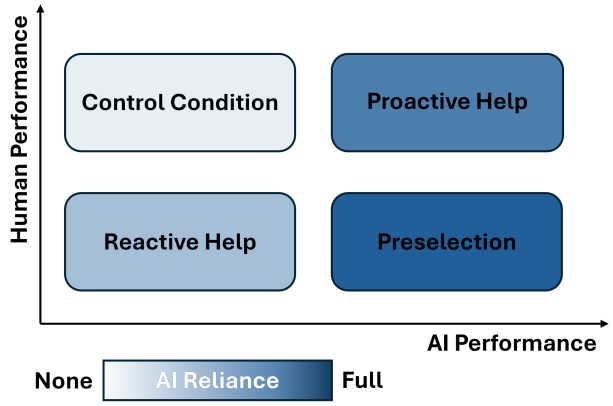

Figure 2. Strategy used for assignment of nudge conditions in the main study.

2.4. Survey Implememtation

For the survey implementation, we used the online tool SoSci hosted on our own servers. The study was preregistered for better transparency before any data was collected [8]. The target audience consisted of participants capable of classifying ECG signals, e.g., medical doctors, medical students, nurses, etc. The survey was divided into a pre-study and a main study. In total, we reached 96 participants, of which 27 took part in the pre-study and 69 in the main study. The participants' exact distribution and professional background are shown in Table 1. Note that eleven participants were excluded from the main study due to one or more of the following reasons: not completing the survey in time, lacking sufficient prior knowledge, or providing meaningless responses.

Table 1. All participants and the exact professions of the main study and all responses.

	All Responses (n=85)		Main Study (n=58)	
Gender	Male	Female	Male	Female
	57	28	45	13
Paramedic	25		25	
Medical Student	10		5	
Medical Staff	16		4	
Other	34		24	

2.4.1. Main Study

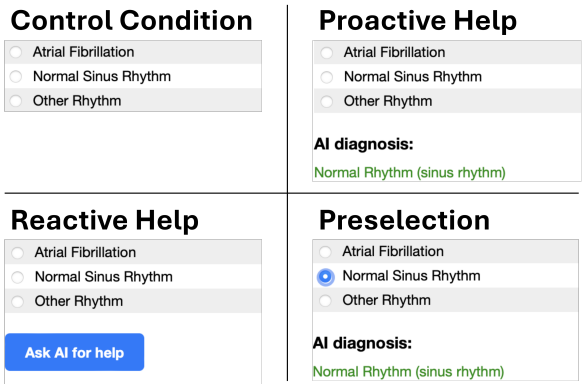

Figure 3. Presentation of nudges in the main study.

For the main study, 58 participants successfully completed the survey. They were evenly divided into a test group and a control group, with the former receiving nudges intelligently and the latter receiving the same nudges but randomly assigned to the tasks. The main study involved the classification of 16 ECGs, selected from the pre-study as relevant cases. The ECG signals were presented as 30-second segments printed on paper to simulate a more realistic clinical setting, similar to what medical staff would encounter in practice. After each classification task, participants were asked to self-report their confidence in their decision-making.

3. Results

The results are presented in two parts. First, we evaluate complementary training as a method to improve AI model performance and support effective AI-human collaboration. Second, we report findings from the online survey, which assessed the feasibility of Intelligent Nudging in improving accuracy.

3.1. Complementary Training

Table 2 summarizes the performance of the complementary and baseline models. As hypothesized, training primarily on difficult samples improved overall performance. The complementary model achieved an F1-score 0.033 higher than the randomly trained model. Interestingly, it also outperformed the baseline model on the easy data subset. These results indicate that, particularly in resource-constrained settings, focusing on difficult samples during training may be beneficial. We attribute the performance gain to the greater variability introduced by our data-splitting strategy.

Table 2. Performance metrics of the models on the easy and difficult test datasets

Performance Metric	Complementary	Baseline
Easy Data		
Accuracy	0.672	0.660
F1-Score	0.679	0.665
Difficult Data		
Accuracy	0.722	0.696
F1-Score	0.712	0.679

3.2. Influence of Intelligent Nudging

To analyze the influence of Intelligent Nudging on overall performance, we compare the results of the four nudging strategies across the intelligent and random groups. The performance results are presented in Table 3. We compare performance across easy, hard, and all data points. The intelligent group outperforms the random group on both easy and hard tasks. Overall, the intelligent group achieves an accuracy that is approximately 20 % higher than that of the random group, a difference that is statistically significant ($p < 0.0035$).

Table 3. Accuracy for the intelligent and random group. The p-value for a one-sided t-test comparing the performance of all collected data points against each other is shown. Values less than 0.05 are considered significant and are highlighted.

	Accuracy		
	Intelligent	**Random**	**p-Value**
Easy Tasks	0.754	0.669	0.0812
Difficult Tasks	0.293	0.203	**0.0479***
All Data	0.524	0.435	**0.0035****

4. Discussion

Our results show that training AI models on human-difficult ECG signals can significantly improve performance on challenging inputs, highlighting the value of complementary training. This approach allows scarce labeling and compute resources to be directed toward cases where AI adds the most value. However, our definition of easy and difficult data was based on handcrafted features. While we consider this sufficient for our work, future work could explore whether AI models themselves can identify samples that are inherently difficult for humans. Our second key finding is that Intelligent Nudging improves human-AI collaboration by aligning AI support with task difficulty. Participants exposed to both Intelligent Nudging performed better than the random nudging group. In conclusion, combining complementary training with Intelligent Nudging offers a promising framework for enhancing decision-making in high-stakes domains like healthcare. Future studies should evaluate these strategies in more diverse populations and real-world clinical environments.

Acknowledgments

We thank Klara Wenzel for conducting the survey implementation and for her analysis of the results. This work was funded by the HEAD-Genuit-Stiftung.

References

[1] Berente N, Gu B, Recker J, Santhanam R. Managing artificial intelligence. MIS quarterly 2021;45(3):1433–1450.

[2] Vishwarupe V, Joshi PM, Mathias N, Maheshwari S, Mhaisalkar S, Pawar V. Explainable ai and interpretable machine learning: A case study in perspective. Procedia Computer Science 2022;204:869–876. ISSN 18770509.

[3] Fügener A, Grahl J, Gupta A, Ketter W. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research 2022;33(2):678–696. ISSN 1047-7047.

[4] Clifford GD, Liu C, Moody B, Lehman LWH, Silva I, Li Q, Johnson AE, Mark RG. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. Computing in Cardiology 2017;44. ISSN 2325-8861.

[5] Hong S, Wu M, Zhou Y, Wang Q, Shang J, Li H, Xie J. Encase: an ensemble classifier for ecg classification using expert features and deep neural networks. Computing in Cardiology 2017;44. ISSN 2325-8861.

[6] Thaler RH, Sunstein CR. Nudge: The final edition. Final edition edition. New Haven: Yale University Press, 2021. ISBN 978-0-241-55210-0.

[7] Goutier M, Diebel C, Adam M, Benlian A. Proactive and reactive help from intelligent agents in identity-relevant tasks. Proceedings of the 57th Hawaii International Conference on System Sciences 2024;.

[8] Goutier M. Intelligent help invocation for improving complementary team performance in logical reasoning tasks. OSF Registries 2024;.

Address for correspondence:

Marc Goutier
Hochschulstraße 1
goutier@ise.tu-darmstadt.de