Rethinking Usability Heuristics for Modern Biomedical Interfaces

Stefan Röhrl*, Christian Janotte*, Christian Klenk†, Dominik Heim†, Manuel Lengl*, Alice Hein*,
Martin Knopp†*, Oliver Hayden† and Klaus Diepold*

*Chair of Data Processing, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany
†Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Einsteinstr. 25, 81675 Munich, Germany

{stefan.roehrl, christian.janotte, christian.klenk, dominik.heim, m.lengl, alice.hein, martin.knopp, oliver.hayden, kldi}@tum.de

***Abstract*—High usability is the ultimate goal in user interface development. In order to test this, user studies are often carried out at great expense. An alternative to this is offered by more favorable implementation guidelines and heuristic evaluation that get by with a smaller number of tests. Tools in the area of biomedical research face major challenges here, as they are extremely crucial, the users are highly demanding, and the advent of Artificial Intelligence (AI) requires researchers to take a powerful leap of faith. Since general heuristics are often insufficient for this domain, we introduce new Biomedical Research AI Heuristics and evaluate them among others using a prototype user interface in the domain of blood cell analysis. The comparative study shows our specialized approach competes very well with Nielsen's well-established general heuristics and a recent publication of rules for AI development. Our set finds the most relevant usability issues and can support the review process for the growing number of biomedical systems that will use artificial intelligence technologies in the future.**

Keywords*—Usability Heuristics; Blood Cell Analysis; Human Assisted Labeling; Quantitative Phase Imaging.

I. Introduction

One of the current challenges in biomedical research is to interpret the increasing amount of data available from new imaging and analysis techniques. To utilize the new information, more and more Artificial Intelligence (AI) is finding its way into this field. It is being used to facilitate differential diagnostics and to improve the understanding of medical conditions. Here, a new platform technology promises major changes in the field of blood analysis. A microscope working with Quantitative Phase Imaging (QPI) does not require expensive reagents and therefore no time-consuming sample preparation [1][2]. Combining this approach with a microfluidics channel, the optical amplitude and phase information of millions of cells can be recorded within minutes. The simplicity, high statistical power and speed of this approach allow statements about the composition of the blood, morphological changes of the cells and thus the kinetics of diseases [3]–[5]. Nevertheless, the resulting images are rather unknown in the medical domain and reference databases as well as sufficient ground truth data is missing, which hinders the efficient training of machine learning algorithms. To overcome these problems, we have to provide an easy way for researchers to work hand in hand with the machine to explore this new field of hematological analysis based on computer vision and AI.

For successful human-computer interaction, the user interface represents the common language the interdisciplinary researchers and developers have to speak. Misunderstandings can prevent such emerging technologies from being successful, as they cannot rely on the trust and the establishment of the gold standard methods [6]. Here, we would like to introduce and compare new **rule set for heuristic evaluation**, which are specifically designed for the development of AI-infused interfaces in biomedical research. As the target group of biomedical researchers and practitioners stands out for a busy schedule and demand high standards in the aspects of explainability [7], transparency [8] and causality [9], having a set of tailor-made heuristics promises a quicker translation of new technologies to the point of care. While most of the usability heuristics used in the past have been of a rather general nature [10], domain-specific ones have become more prominent in the last decades [11]–[13].

In this work, we propose a new labeling platform for holographic cell images where humans and AI work closely together in (inter-)active learning scenarios. This will facilitate the generation of verified ground truth data and be a valuable representative for this kind of biomedical user interfaces. Our primary interest, however, is to validate the newly developed usability heuristics against the existing ones, and thus to meet the need for guidance in the development process of AI-infused biomedical systems.

In the following, the work is divided into the appropriate sections: Section II motivates the choice of the clinical application and introduces the concepts for comparing heuristic rule frameworks. Then, Section III presents the specially developed web-based prototype of a user interface. The three sets of heuristic rules are introduced in Section IV, followed by their evaluation by experts as well as by user tests in Section V. The results of the study are described and visualized in Section VI. Finally, Section VII discusses the findings and draws conclusions for possible future work.

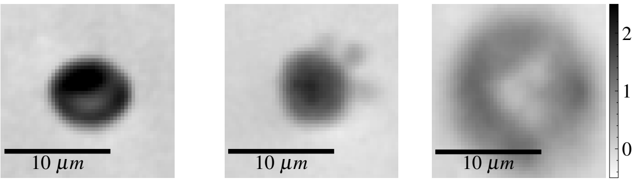
II. Background and Related Work

Before introducing the prototype, we will investigate the medical relevance of the chosen use case and the methods for evaluating sets of heuristic rules.

A. Medical Relevance of Quantitative Phase Imaging

The process of blood analysis in general is one of the most requested laboratory tests [14] and has been extensively studied in the past, leading to technically advanced solutions.

Copyright (c) IARIA, 2023. ISBN: 978-1-68558-078-0



(a) Red Blood Cell (b) Aggregate (c) Defocused Cell

Figure 1. Phase images of different cell classes

As a result, most state-of-the-art instruments work with a blood processing scheme based on marker materials [6]. Although these devices being very precise, they come with several downsides, as they require non-specific and costly labeling as well as time-consuming sample preparation such as *hemolysis* [15]. Using QPI methods combined with machine learning, the exercise translates into a computer vision task, which offers more flexibility. The morphological and internal patterns of blood cells provide insights for oncological [3][16], parasitic infections [17] and other diseases [4]. Also, the aggregation of blood cells can deliver crucial information [5][18].

However, before the images can be automatically interpreted and classified, they must be segmented and labeled by experts. Figure 1 shows representatives of typical cells and structures as they look like under a quantitative phase microscope. Red blood cells (a) are quite simple to detect, whereas aggregates of white blood cells and platelets (b) are more difficult to find due to their complex structure and the associated rarer occurrence. The algorithm as well as the human also have to learn, which objects need to be discarded (c). Note that medical experts are usually only trained on stained thin films and are therefore unfamiliar to this representation [6]. The brightness information directly correlates with the optical phase shift $\Delta\phi$ caused by the cells. Greater detail about the microscope can be found in [2][3].

B. Active Learning for Human Assisted Labeling

Manually labeling large amounts of data such as images is tedious and sometimes even challenging for skilled personnel, as the previous section describes. Therefore, crowd sourcing is not an option. As biomedical experts are expensive and limited in time, the Active Learning (AL) approach seems promising [19]. In AL, an algorithm is trained on a very sparse data set to learn a classification problem. However, instead of leaving the user with the task of correcting a predicted class label when the system is uncertain, the algorithm attempts to minimize the actions that need to be taken [20]. Moreover, AL shows suitable behavior for imbalanced data sets like ours to build a *human-in-the-loop* system [21], as we do in our prototype.

C. Quality Assessment of Usability Heuristics

The developed user interface represents the precedent to put our newly developed heuristics into practice. To make the heuristics more comparable, we need to introduce quality assessment measures as well as standard procedures to obtain these measures. Hartson et al. [22] propose to apply the different evaluation methods to the target system and compare the found usability problems to a baseline of "real" usability problems. In our work, we will determine the baseline by conducting *asymptotic user testing* [22]. As not every usability problem is as crucial as the other, we will further rate each problem then by a *severity score* proposed by Nielsen [23]. Table I shows the weighting of the apparent usability problems in order to compare the heuristics on their ability to prevent major usability issues.

TABLE I. SEVERITY RATINGS FOR USABILITY PROBLEMS [23]

	$s(p)$	**Description**
Rating	0	Violates a heuristic but is not a usability problem
	1	Cosmetic or unimportant usability problem
	2	Minor usability problem
	3	Significant usability problem
	4	Usability catastrophe

Starting from there, Sears [24] defines the **thoroughness** criterion (also known as recall in other disciplines)

$$T = \frac{|E \cap F|}{|E|}, \tag{1}$$

where $|E \cap F|$ denotes the number of problems F found by the heuristics from the baseline set of real usability problems E. Using our mapping of severity scores we can calculate the **weighted thoroughness**

$$T_w = \frac{\sum_i s(f_i)}{\sum_j s(e_j)} \text{ with } f_i \in E \cap F \text{ and } e_j \in E, \tag{2}$$

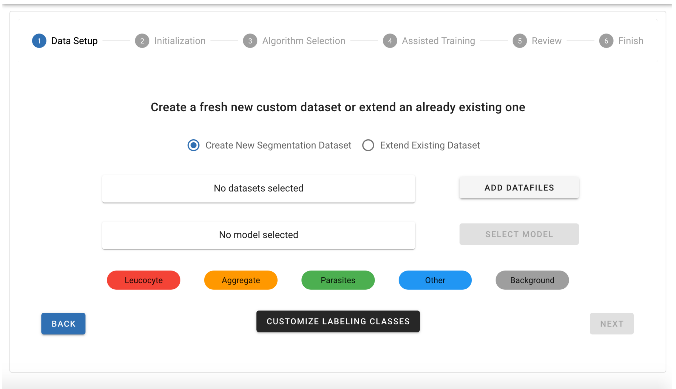
where $s(p)$ assigns every usability problem its rating according to Table I. Finally, the **validity** criterion [25] (also called precision)

$$V = \frac{|E \cap F|}{|F|} \tag{3}$$

helps us to judge how many of the identified problems F where real and no false alarms.

III. HUMAN ASSISTED LABELING PROTOTYPE

In order to provide an easily accessible and customizable user interface, we developed a web-based prototype for this study, which is divided into different views.



Create a fresh new custom dataset or extend an already existing one

● Create New Segmentation Dataset ○ Extend Existing Dataset

No datasets selected

No model selected

Figure 2. View 1 - Data Setup

The **data setup** view displayed in Figure 2 contains the functionality for setting up the general properties of the project. At the top, there is an option button that allows the user to choose whether to start with an empty data set or expand an existing data set based on a previously trained algorithm. Below, the user finds means to load the respective data containers or models. The lower part of the page displays the currently available classes of cell types. Each of them has its own color scheme and can be customized, added or deleted by clicking the button below them.

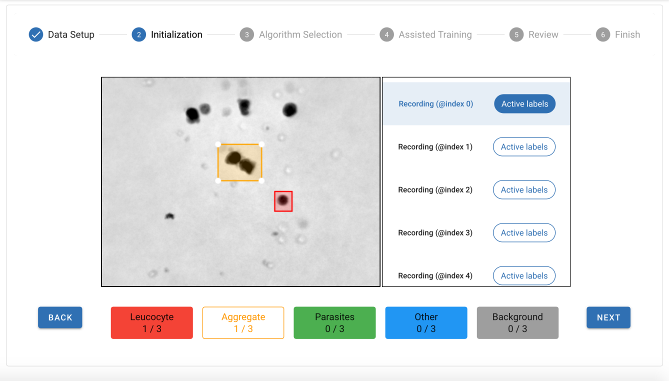
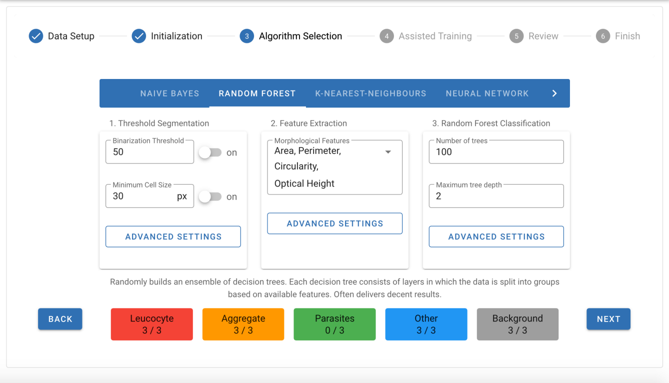


Figure 3. View 2 - Initialization

After having specified the labeling task and the expected classes of cells, clicking the "Next" button opens the **initialization** view (Figure 3). As the name suggests, it is used to provide an initial training set for the later algorithm. A large canvas is the main component of this view, displaying the selected set of cells, but also providing an area for drawing and annotating. In the bottom part of the view, there is a footer that displays the available classes. Clicking on one of them activates the class which is illustrated by highlighting. The user can now click and drag the mouse to draw bounding rectangles around the cells in the image. This combination of location and class is later called a label.



1. Threshold Segmentation

Binarization Threshold: 50 — on

Minimum Cell Size: 30 px — on

2. Feature Extraction

Morphological Features: Area, Perimeter, Circularity, Optical Height

3. Random Forest Classification

Number of trees: 100

Maximum tree depth: 2

Randomly builds an ensemble of decision trees. Each decision tree consists of layers in which the data is split into groups based on available features. Often delivers decent results.

Figure 4. View 3 - Algorithm Selection

In the **algorithm selection** view (Figure 4), users can specify the type of algorithm they want to use to classify cells in the records by selecting the appropriate tab at the top. Currently, users can choose from *Naive Bayes*, *Random Forest*, *k-Nearest-Neighbors* and a small *Neural Network* [26][27]. Depending on the type of classifier, necessary segmentation and feature extraction steps can be customized in the respective tab.

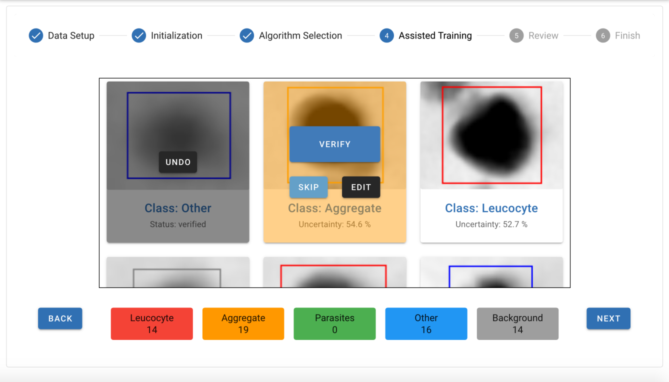
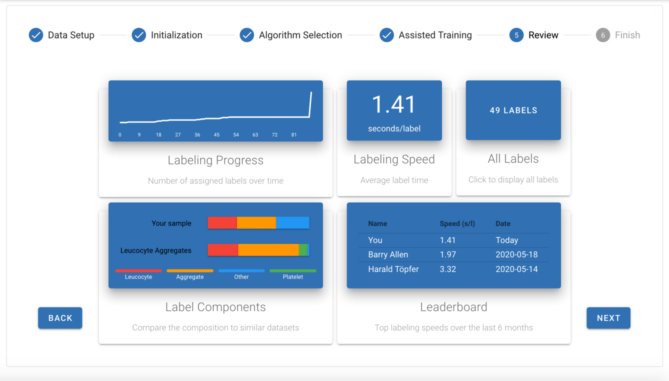


Figure 5. View 4 - Assisted Training

When the algorithm has made its first predictions based on the initial training set the **assisted training** part starts in this reoccurring view (Figure 5). A gallery appears showing the proposed labels that the algorithm found in the data. As suggested by the AL principles from Section II-B, they are ordered by their uncertainty from the highest to the lowest value. Here, users can intervene and verify or correct the algorithm and hence, enlarge the training set without manually scanning the raw data and drawing rectangles. Furthermore, human assistance is only required for difficult objects, reducing the wasted time on already mastered samples. The algorithm can then be periodically retrained on the extended training set and can quickly reach a satisfying performance on the complete data set.



Name	Speed (s/l)	Date
You	1.41	Today
Barry Allen	1.97	2020-05-18
Harald Töpfer	3.32	2020-05-14

Figure 6. View 5 - Review

Finally, the **review** view (Figure 6) summarizes the labeling progress over time and the current performance. It compares the composition of the data set to other similar data sets and displays the percentages of detected cell classes. As a kind of gamification element, it also shows the labeling speed and ranks it with the performance of other users.

ISBN: 978-1-68558-078-0

IV. Usability Heuristics

This section gives an overview of the state of the art in usability heuristics and introduces our new set of rules.

A. Nielsen's Heuristics

The general usability heuristics by Nielsen and Molich have been known for decades and are still used today. They are based on some of the most fundamental rules for user interface development. Their strength in finding many usability problems has been demonstrated in the past. Due to their generality, they can be applied to almost any type of system, but they have the disadvantage of not always finding as many usability problems as a set developed specifically for the system's domain. Nevertheless, they are a good starting point and will be a strong competitor and thus valuable for comparing them with our own set of heuristics. For our comparison, we used the rules from Nielsen's most cited publications [10], [28]. Also, minor modifications in wording [29], done in the last years, were considered.

B. Human-AI Interaction Heuristics

With the advent of AI in recent decades, it was only a matter of time before user interface developers began to address the specific requirements of AI-infused interfaces. In collaboration with Microsoft, a group of researchers led by Amershi recently proposed a set of usability guidelines for the development of such systems [30]. Guidelines and heuristics are not technically the same thing, since guidelines are used during the implementation of an interface and heuristics during verification. However, for short lists of guidelines such as this one, they can often be used interchangeably [31]. For our experiments, we converted the guidelines into a set of heuristic rules and provided them with examples for the experts so that this set can be used equivalently for the evaluation. This set of heuristics additionally distinguishes whether a problem occurs **immediately**, while the user is using the tool, or if it appears over a **longer period**. It is to be expected that some of the listed rules of this set will have minor relevance for our labeling interface like Rule 6 that is about mitigating social biases. To be consistent, we will keep the set unaltered.

C. Biomedical Research AI Heuristics

The main idea of this work is not only to compare the proven heuristics by Nielsen and Molich and the recently published AI guidelines interpreted as heuristics. We intend to create our own set of heuristics specifically targeted at biomedical research applications that use AI. The amount of software in this area will increase in the coming years, and it may be beneficial to have custom heuristics at hand for evaluation to save valuable testing time. Table II shows a set of 15 rules grouped in four categories, which constitute our *Biomedical Research AI Heuristics*. They are inspired by several publications in the domain of user interface design, biomedical and AI applications over the last decades. We completed those rules by hints and suggestions from preceding interviews with experts from local institutions working in the field of biomedical research.

Table II. Heuristics for AI in biomedical research

	#	Name	Short Description
Structure	1	Streamline main task	Focus on the main task that a system was created for and make the system easy to learn [32].
	2	Provide full control	Provide global control of important model parameters and the data pipeline [33][34].
	3	Orientation	Always show users where they are, what is currently going on and what they can do next [10].
Interaction	4	Guide attention	Keep the users focused on their task and only alarm them in urgent cases [35][36].
	5	Provide comparisons	Let users compare among similar data or parameters when they need to judge an outcome or make a decision.
	6	Show impact	Users need to see how their actions influence the system and its performance [37].
	7	User over System	Allow users to correct errors of the AI efficiently at all times and even turn off the AI if needed [35].
Presentation	8	Familiar language	Use non-technical language if possible. Pay attention to use correct terminology for medical concepts [38].
	9	Precise language	Avoid ambiguous wording for labels and commands that could trigger confusion [10].
	10	Familiar look	Use ways of presentation for the interface that users know from other tools.
	11	Appeal	Give the users the feeling of using a state-of-the-art and high-quality product.
Explainability	12	Explain data	Foster the interpretability of the data and how it differs from other data sources [39].
	13	Explain processing	There needs to be a high-level explanation for the overall procedure that is performed by the system [9].
	14	Explain reasoning	There has to be an explanation why and how the system derived a certain result or prediction [9].
	15	Strengths / Limitations	Show what the strengths and weaknesses of the system are and what expectations are realistic. [40]

V. Usability Evaluation

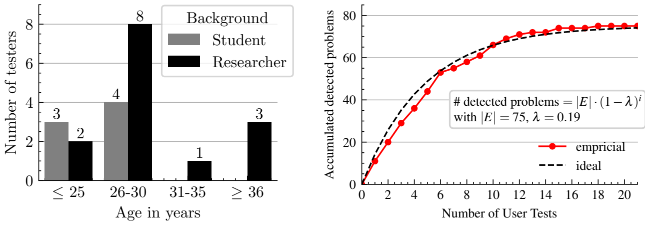
Once all the prerequisites are met, the prototype is tested by means of heuristic evaluation and user testing.

A. Heuristic Evaluation

For the evaluation of heuristics, we will compare the three heuristics with different aims and origins presented in the previous section. Their performance will be compared to determine whether general or domain-specific heuristics perform better in the domain of AI-infused interfaces for biomedical research. Most usability researchers like Nielsen classify potential expert evaluators into three different categories: *novices*, *single experts* and *double experts* [41]. Novices are new to usability concepts but often have knowledge in the domain where the user interface will be deployed. In contrast, single experts already have experience in the field of usability engineering but lack knowledge of the designated domain. Double experts are evaluators who are proficient both in usability engineering and the domain. On average, a novice finds only 22% of issues in a system, while single experts manage to find 41% and double experts even around 60% [42]. The experts participating in our review are neither novice evaluators nor have they been conducting such reviews for years. Nevertheless, they have a sound knowledge of usability concepts and have conducted a heuristic evaluation before. In addition, some of them also have a basic understanding of the domain of the system. Each heuristic is applied to our prototype user interface by five different evaluators, a number often recommended for user interface development because of its cost-efficiency [43]. In order to keep focus on the most relevant usability problems, we use the severity rating system introduced in Table I. During the expert review process, each expert will assign a level of potential impact to the usability problems they have discovered. After a final list of aggregated usability problems is compiled for all heuristics, each expert will also assign ratings to the problems found by their peers. In the end, the ratings among the experts will be averaged and rounded.

B. User Testing

In order to compare the different heuristics in this work, we need to gather knowledge about the real usability problems E inherent in our prototype interface. For this, asymptotic user testing [22] is selected as a test procedure. With a conservative detection rate of 19% per user [22][44], the relation between the number of testers and the percentage of discovered usability problems seems to level off at around 20 testers, which is very late. This is shown by the ideal curve in Figure 7b. However, to increase the chances of overlooking as few problems as possible, we decided to conduct a test series with at least this number of testers. Eventually, we found 21 representative users with a biomedical background who were willing to participate. Their demographics are displayed in Figure 7a. The youngest tester was 21 and the oldest 59 years old. What almost all testers had in common was their lack of experience with machine learning. 76.2% said they had no experience at all. This was beneficial to see how they would react to something they had never used before.

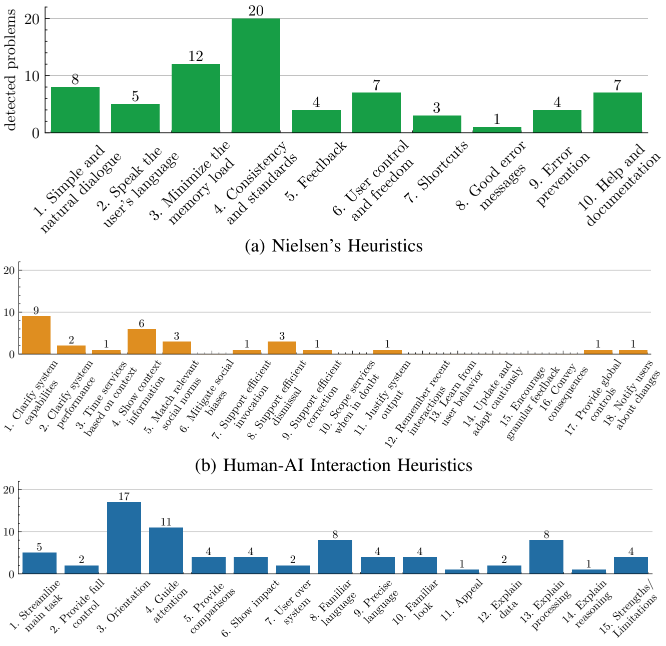


(a) User Demographics

(b) Detected problems by users

Figure 7. Composition and performance of the user tests

All users were given two tasks: 1) "In a small sample you are interested in the number of white blood cells, single platelets, and cell aggregates. Extract these components and perform some further evaluation to show them to a co-worker." 2) "Your bigger recording is rich in white blood cell aggregates. You want to detect the same components as before but also keep track of other cells as they might become relevant later. Prepare and store your results for further evaluation." Users were given as much time as they needed to complete the tasks and were encouraged to ask questions and think aloud throughout the test [45]. Meanwhile, the evaluator took informal notes that would later be summarized in a formal test protocol. Testers were also required to complete a short questionnaire after the test.



(a) Nielsen's Heuristics

(b) Human-AI Interaction Heuristics

(c) Biomedical Research AI Heuristics

Figure 8. Detected problems by the respective heuristics

VI. Results

This section summarizes findings and compares the results of the different testing strategies by the proposed metrics.

A. Heuristic Evaluation

The first set of rules we applied to our biomedical user interface consisted of **Nielsen's ten general usability heuristics**. They were developed without regard to a specific type of user interface. The review conducted by five evaluators using this rule set identified 60 violations within the system. This number was obtained by comparing and aggregating the results of the individual evaluators. Figure 8a shows the number of usability problems identified by each rule of Nielsen's heuristics. It is important to note that the sum of all bars is greater than the total number of problems identified, since a problem may relate to more than one heuristic. Prominently, Rule 4, which deals with user interface consistency and standards, is responsible for 20 usability problems, which is significantly more than any other rule. The second most problems are related to Rule 3, which focuses on reducing the user's memory load. It is not possible to say whether their numerous occurrence is due to the fact that these rules highlight important aspects of biomedical interfaces very effectively, or whether an unusual high number of violations occurred by chance. The rough list of usability problems merely indicates the presence of these violations. Conclusively, the five evaluators emphasized that they enjoyed working with the set and that it was easy to use. In addition, it is worth noting that each heuristic was applied at least once and no heuristic was omitted.

The second set of rules used in this project were the recently published **guidelines for human-AI interaction**. They were proposed as guidelines that can support the development of interfaces to let human users interact with AI. In our evaluation, the five experts discovered 26 violations and the corresponding usability problems, which is less than half the amount that Nielsen's heuristics helped to find. Figure 8b shows the number of heuristic violations per rule in this set of human-AI heuristics. The distribution of problems looks quite different from that resulting from Nielsen's heuristics. First, there are a number of rules that did not help uncover a usability problem at all. This is mainly due to the fact that these aim for long-term effects which do not apply to the tasks covered in our study. The two heuristics that have received the most attention are Rule 1, which deals with explaining what the system can do, and Rule 4, concerning context and relevancy of the displayed information. What is interesting about this second set of heuristics is the informal feedback from the evaluators. They pointed out that these rules were very difficult to apply to the system. The reason for this could be that they were not developed as heuristics, but as guidelines. As such, they might be too specific and not generally applicable.

The third set of rules we applied to the interface is the one we created specifically for the field of **biomedical research interfaces that use AI**. Here, the five experts reported a list of 55 usability problems. This is slightly less than what they discovered with the general heuristics, but still much more than what the heuristics for human-AI interaction identified. The distribution of usability issues across the different rules within our custom heuristics is shown in Figure 8c. All fifteen rules were found to have at least one violation. The two most frequent heuristics are Rule 3 and Rule 4, which are concerned with providing orientation and guiding the user's attention. The third place is shared by Rule 8 and Rule 13. It is interesting to note that these four heuristics are all aimed at reducing the complexity of AI for the biomedical users or enabling them to better deal with it. Evaluators noted that the set was easy to use and that they felt it covered most usability issues with a large impact on the user experience. This feeling is supported by the fact that it detected the most usability issues with the highest impact among the three heuristic sets, with fourteen violations of the maximum severity level.

B. User Testing

This would lead us to the quality assessment metrics introduced in Section II-C, but before we can apply them we need the baseline of real usability problems E determined by our user tests. With respect to the asymptotic behavior of the usability problem discovery process, we assumed that about 20 testers would be needed to find most of the problems. The test ultimately resulted in the detection of 75 usability problems over the course of 21 user tests. To support the claim that we almost reached an asymptotic upper bound, we plotted the occurrence of problems over tests in Figure 7b, indeed revealing the asymptotic shape of a Poisson process [43].

To obtain the severity ratings of the real usability problems, we sent the complete list of issues to our usability experts and summarized the ratings based on their judgment. Many of the entries in this list are common problems that can occur in any type of user interface, such as misleading button descriptions and lack of loading indicators. However, there are also some problems (see Table III) that seem to be rather unique and that can serve as examples of typical problems in environments where users with a biomedical background interact with AI. These were concentrated to uncertainties about the specific workflow of the program and obscure consequences, which certain changes in the settings might have. Only 2 out of 21 users requested major changes before they would use such a system for their daily work. 19% stated that they would use it, but still suggested some minor changes. The majority of 71% of users indicated that they would use the system in the future exactly as it is, after becoming familiar with it.

TABLE III. EXEMPLARY USABILITY PROBLEMS

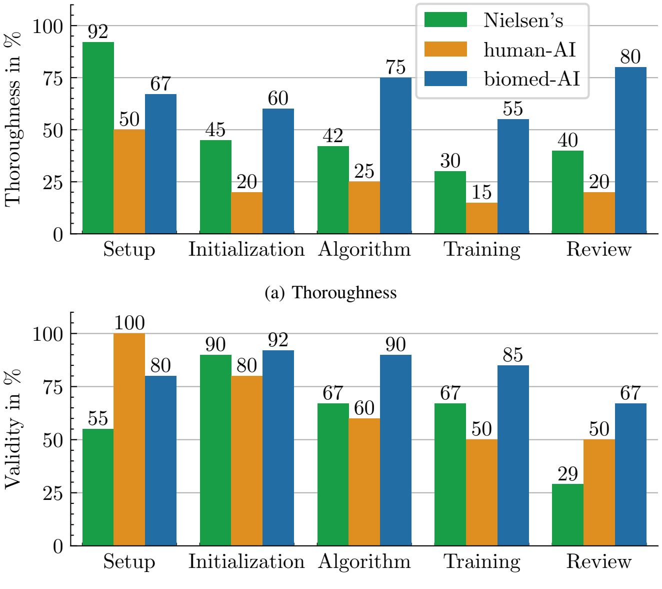
View	**Description** **Note:** The listed problems all have a maximum severity rating of 4. The numbers indicate the violated heuristic rule or the number of affected users respectively.	Nielsen's	human-AI	biomed-AI	User Test
1	There is no clear indicator that tells the user when the initialization is completed or what happens with empty classes. The "X/3" in the footer is not prominent enough.	5 10		4 6	2
3	The different algorithms are not sufficiently explained and the current explanations are hard to find. Users do not know which algorithm to choose.	1 2		4 6 13 15	1
3	The wording of some parameters and explanations is too technical to understand.			8	2
4	Users do not understand the training process, what they have to do and why multiple iterations with retraining make sense. The initial performance might be disappointing.	3 10	1 2	13	5

C. Metric-Based Comparison of Heuristics

Now that we have a baseline, we can relate it to the findings from different heuristics. This results in a list of 104 usability problems, with which we can compute the quality assessment metrics. As listed in Table IV, the three different sets of heuristics did not perform equally well. For almost all metrics, the domain focus of our set of heuristics is noticeable and provides improved results in the criteria **thoroughness** and **validity**. The general heuristics by Nielsen still occupy a stable second place, although it should be noted that all three heuristics were not able to predict usability problems seamlessly. Nevertheless, the high validity of our custom heuristics make them a reliable tool to alert developers of incipient and severe usability issues. We can further compute the thoroughness metric for high severity levels (3 & 4), as these should be addressed early in the development process. Among the highest level of severity (4), our biomedical heuristics account for a thoroughness of

TABLE IV. RESULTS OF THE QUALITY ASSESSMENT METRICS

Metric	**Nielsen's**	**human-AI**	**biomed-AI**
Thoroughness	50.1%	25.3%	**62.7%**
Weighted Thoroughness	54.0%	28.9%	**69.0%**
Average Severity	2.66	**2.84**	2.74
Validity	63.3%	73.1%	**85.5%**



(a) Thoroughness

(b) Validity

Figure 9. Quality assessment metrics for the individual views

93.3%. Nielsen's set improves to 73.3% and the human-AI interaction heuristics to 46.7%.

Since not all views of our prototype are equally influenced by AI, we investigate the performance in the individual components of the system. Accordingly, Figure 9a shows the **thoroughness** metric for each set of heuristics. The Nielsen heuristics have the highest thoroughness in the setup view, but as more interaction with the AI is emerging, their performance drops. Surprisingly, the AI emphasized human-AI heuristics score even worse. Our biomedical heuristics have the highest thoroughness in the initialization, algorithm, training and review views, as these require frequent interaction between the users and the machine learning algorithms.

Similarly, we can evaluate the **validity** of the heuristics depending on the view as displayed in Figure 9b. The validity of our biomedical heuristics is always higher as Nielsen's heuristics. This means that Nielsen's heuristics tend to find a lot of irrelevant usability issues in all views. However, the validity for the review view is particularly low, for all three sets.

VII. Discussion and Conclusion

The performance metrics from Section VI-C indicate that there is a noticeable difference in which heuristics we use for an expert evaluation of an AI-infused user interface within the biomedical domain.

Nielsen's well-known heuristics struggle when it comes to finding real usability problems in biomedical interfaces induced with AI. They showed only mediocre thoroughness in these parts of the prototype. However, they found the most genuine usability problems in the parts of the interface that were least affected by machine learning, resulting in a high performance in those views. Unfortunately, this seems to be accompanied by reduced validity. Nielsen's heuristics tend to find more expendable problems than the competing heuristics. All in all, the results suggest that these general heuristics are not always the best choice when it comes to finding usability problems in a specific domain like the one we studied. This is a result that also has been discussed in other publications [46].

The **heuristics for human-AI** interaction did not score particularly well in terms of thoroughness and validity. In addition, the experts in this study indicated that this set was most difficult to use for interface evaluation. This could be due to the fact that this set was originally designed as a guideline and also has large focus on long-term effects that are not relevant here.

The **heuristics for biomedical user-AI** interaction that we developed in this work provided the most compelling results. While their thoroughness was good but not great, their weighted thoroughness and thus their potential to uncover the most important problems in a user interface like our prototype was a positive discovery. This was further emphasized by the set's high thoroughness scores for high severity problems. Moreover, our set performed better than Nielsen's general heuristics, especially in the parts of the interface that focused on user-AI interaction.

When putting the heuristics' evaluation in a larger context, we expected that we could detect at least 70% of the real usability problems as foreseen in literature [11][22][43]. Our experts were not novices, but the best detection rate they could achieve was 62.7% with the biomedical heuristics and even less with the other sets. There is a possibility that this is due to inadequate evaluation of our experts. However, it is more likely that the main reason is that it is simply more difficult to find usability problems in the domain we analyzed. This assertion is supported by studies like [11], pointing out the need for domain-specific heuristics for domains where usability problems are immanently difficult to detect. This was also one of the basic assumptions on which this entire paper is based. As biomedical interfaces seem challenging, an unweighted thoroughness of 62.7% is a satisfactory result.

Finally, we aim to apply our new biomedical heuristics on more user interfaces in this domain. Tools that are used for making diagnoses and more complex reasoning could be of special interest. With a more diverse expert group, we hope to reduce the effort of conducting user tests and help to establish AI based technologies in biomedical research and healthcare.

Acknowledgment

The authors would like to especially honor the contributions C. Janotte for the implementation of the user interface prototype and the execution of the studies. Furthermore, thanks go to D. Heim and C. Klenk for the sample preparation and recording of the measurements. All expert reviewers should also be thanked at this point.

This research was funded by the German Federal Ministry for Education and Research (BMBF) with the funding ID ZN 01 | S17049.

References

[1] Y. Park, C. Depeursinge, and G. Popescu, "Quantitative phase imaging in biomedicine," *Nature Photonics*, vol. 12, no. 10, pp. 578–589, 2018.

[2] C. Klenk, D. Heim, M. Ugele, and O. Hayden, "Impact of sample preparation on holographic imaging of leukocytes," *Optical Engineering*, vol. 59, no. 10, p. 102403, 2019.

[3] M. Ugele, M. Weniger, M. Stanzel, M. Bassler, S. W. Krause, O. Friedrich, O. Hayden, and L. Richter, "Label-Free High-Throughput Leukemia Detection by Holographic Microscopy," *Advanced Science*, vol. 5, no. 12, 2018.

[4] T. L. Nguyen, S. Pradeep, R. L. Judson-Torres, J. Reed, M. A. Teitell, and T. A. Zangle, "Quantitative Phase Imaging: Recent Advances and Expanding Potential in Biomedicine," *American Chemical Society Nano*, vol. 16, no. 8, pp. 11 516–11 544, 2022.

[5] M. Nishikawa, H. Kanno, Y. Zhou, T.-H. Xiao, T. Suzuki, Y. Ibayashi, J. Harmon, S. Takizawa, K. Hiramatsu, N. Nitta *et al.*, "Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with covid-19," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021.

[6] J. J. Barcia, "The Giemsa stain: Its History and Applications," *International Journal of Surgical Pathology*, vol. 15, no. 3, pp. 292–296, 2007.

[7] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 359–380.

[8] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mandl, "Machine intelligence in healthcare - perspectives on trustworthiness, explainability, usability, and transparency," *Nature Partner Journals Digital Medicine*, vol. 3, no. 1, pp. 1–5, 2020.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[10] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 249–256.

[11] S. Hermawati and G. Lawson, "Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus?" *Applied Ergonomics*, vol. 56, pp. 34–51, 2016.

[12] A. W. Kushniruk, V. L. Patel, and J. J. Cimino, "Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces." in *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, 1997, p. 218.

[13] J. Zhang, T. R. Johnson, V. L. Patel, D. L. Paige, and T. Kubose, "Using usability heuristics to evaluate patient safety of medical devices," *Journal of Biomedical Informatics*, vol. 36, no. 1-2, pp. 23–30, 2003.

[14] S. Horton, K. A. Fleming, M. Kuti, L.-M. Looi, S. A. Pai, S. Sayed, and M. L. Wilson, "The Top 25 Laboratory Tests by Volume and Revenue in Five Different Countries," *American Journal of Clinical Pathology*, vol. 151, no. 5, pp. 446–451, 2018.

[15] A. Filby, "Sample preparation for flow cytometry benefits from some lateral thinking," *Cytometry Part A*, vol. 89, no. 12, pp. 1054–1056, 2016.

[16] S. K. Paidi, P. Raj, R. Bordett, C. Zhang, S. H. Karandikar, R. Pandey, and I. Barman, "Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia," *Biosensors and Bioelectronics*, vol. 190, p. 113403, 2021.

[17] M. Ugele, M. Weniger, M. Leidenberger, Y. Huang, M. Bassler, O. Friedrich, B. Kappes, O. Hayden, and L. Richter, "Label-free, high-throughput detection of P. falciparum infection in sphered erythrocytes with digital holographic microscopy," *Lab on a Chip*, vol. 18, pp. 1704–1712, 2018.

[18] M. Finsterbusch, W. C. Schrottmaier, J. B. Kral-Pointner, M. Salzmann, and A. Assinger, "Measuring and interpreting platelet-leukocyte aggregates," *Platelets*, vol. 29, no. 7, pp. 677–685, 2018.

[19] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[20] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[21] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[22] H. R. Hartson, T. S. Andre, and R. C. Williges, "Criteria for evaluating usability evaluation methods," *International Journal of Human-Computer Interaction*, vol. 13, no. 4, pp. 373–410, 2001.

[23] J. Nielsen, "Severity ratings for usability problems," *Papers and Essays*, vol. 54, pp. 1–2, 1995.

[24] A. Sears, "Heuristic walkthroughs: Finding the problems without the noise," *International Journal of Human–Computer Interaction*, vol. 9, no. 3, pp. 213–234, 1997.

[25] W. D. Gray and M. C. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods," *Human Computer Interaction*, vol. 13, no. 3, pp. 203–261, 1998.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[28] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994, pp. 152–158.

[29] ——. (2005) Ten usability heuristics. (accessed 2022.12.17). [Online]. Available: https://www.informaticathomas.nl/heuristicsNielsen.pdf

[30] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[31] B. Shneiderman, *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley, 1998.

[32] H. Lieberman, "User interface goals, AI opportunities," *AI Magazine*, vol. 30, no. 4, pp. 16–22, 2009.

[33] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[34] T. G. Gill, "Expert systems usage: Task change and intrinsic motivation," *Management Information Systems Quarterly*, pp. 301–329, 1996.

[35] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[36] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *Nature Partner Journals Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[37] A. D. Jameson, "Understanding and dealing with usability side effects of intelligent processing," *AI Magazine*, vol. 30, no. 4, pp. 23–23, 2009.

[38] C. Rzepka and B. Berger, "User interaction with AI-enabled systems: a systematic review of is research," in *Thirty Ninth International Conference on Information Systems*, vol. 39, 2018, pp. 1–17.

[39] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[40] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019.

[41] J. Nielsen, "Finding usability problems through heuristic evaluation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1992, pp. 373–380.

[42] J. Noyes and C. Baber, *User-centred design of systems*. Springer Science & Business Media, 1999.

[43] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, 1993, pp. 206–213.

[44] J. R. Lewis, "Sample sizes for usability studies: Additional considerations," *Human factors*, vol. 36, no. 2, pp. 368–378, 1994.

[45] M. W. Jaspers, T. Steen, C. Van Den Bos, and M. Geenen, "The think aloud method: a guide to user interface design," *International Journal of Medical Informatics*, vol. 73, no. 11-12, pp. 781–795, 2004.

[46] C. Jimenez, P. Lozada, and P. Rosas, "Usability heuristics: A systematic review," in *2016 IEEE 11th Colombian Computing Conference (CCC)*. IEEE, 2016, pp. 1–8.