"If it is easy to understand, then it will have value": Examining Perceptions of Explainable AI with Community Health Workers in Rural India

CHINASA T. OKOLO*, Cornell University, United States
DHRUV AGARWAL*, Cornell University, United States
NICOLA DELL, Cornell Tech, United States
ADITYA VASHISTHA, Cornell University, United States

AI-driven tools are increasingly deployed to support low-skilled community health workers (CHWs) in hard-to-reach communities in the Global South. This paper examines how CHWs in rural India engage with and perceive AI explanations and how we might design explainable AI (XAI) interfaces that are more understandable to them. We conducted semi-structured interviews with CHWs who interacted with a design probe to predict neonatal jaundice in which AI recommendations are accompanied by explanations. We (1) identify how CHWs interpreted AI predictions and the associated explanations, (2) unpack the benefits and pitfalls they perceived of the explanations, and (3) detail how different design elements of the explanations impacted their AI understanding. Our findings demonstrate that while CHWs struggled to understand the AI explanations, they nevertheless expressed a strong preference for the explanations to be integrated into AI-driven tools and perceived several benefits of the explanations, such as helping CHWs learn new skills and improved patient trust in AI tools and in CHWs. We conclude by discussing what elements of AI need to be made explainable to novice AI users like CHWs and outline concrete design recommendations to improve the utility of XAI for novice AI users in non-Western contexts.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning approaches**;

Additional Key Words and Phrases: Artificial Intelligence, Machine Learning, Community Health Workers, Mobile Health, Explainability, HCI4D, XAI4D, ICTD, Global South

ACM Reference Format:
Chinasa T. Okolo, Dhruv Agarwal, Nicola Dell, and Aditya Vashistha. 2024. "If it is easy to understand, then it will have value": Examining Perceptions of Explainable AI with Community Health Workers in Rural India. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 71 (April 2024), 22 pages. https://doi.org/10.1145/3637348

*Both authors contributed equally to this research.

Authors' addresses: Chinasa T. Okolo, Computer Science, Cornell University, 350 Gates Hall, Ithaca, New York, United States, chinasa@cs.cornell.edu; Dhruv Agarwal, Information Science, Cornell University, Ithaca, New York, United States, da399@cornell.edu; Nicola Dell, Information Science, Cornell Tech, New York, New York, United States, nixdell@cornell.edu; Aditya Vashistha, Information Science, Cornell University, Ithaca, New York, United States, adityav@cornell.edu.

2573-0142/2024/4-ART71 $15.00
https://doi.org/10.1145/3637348

1 INTRODUCTION

Many low-income countries in the Global South have established community health programs that rely on paraprofessional community health workers (CHWs) to provide last-mile health services [? ?]. CHWs are usually women with a high school education who are recruited from local communities and receive a few weeks of medical training to provide essential health services according to approved health protocols [? ?]. Although CHWs have improved health outcomes, such as increased uptake of breastfeeding, immunizations, and institutional deliveries [? ?], their performance has been suboptimal largely due to their low levels of medical knowledge [?]. In lieu of upskilling opportunities, AI-driven tools are rapidly integrated into their workflows to help them diagnose diseases [? ?], analyze rapid diagnostic tests [? ?], and manage patient-care [? ? ?]. However, crucial for a fruitful CHW-AI collaboration is the ability of CHWs to understand the results of the models powering these tools.

A large body of scholarship on explainable AI (XAI) thus far has focused on Western contexts or on specialized practitioners such as pathologists [?], which are remarkably different from community health care contexts that comprise socially, culturally, and technologically diverse CHWs in the Global South. Although a nascent body of HCI scholarship has begun to explore how concepts of AI fairness and explainability differ across Western and non-Western contexts [? ?], little is known about what CHWs, who possess low AI literacy and operate AI-driven tools in high-stakes settings, need to know to cooperatively work with AI-driven tools and become effective AI workers. To fill this critical gap, we sought to answer the following research questions:

RQ1: How do CHWs engage with and perceive AI explanations?
RQ2: How do we design XAI interfaces that are more understandable to them?

To answer these questions, we conducted semi-structured interviews with 35 CHWs who interacted with a design probe that predicts neonatal jaundice and provides explanations on how it arrived at a prediction. Given the low levels of AI literacy in CHWs, we used design probes because they have been shown to be effective in eliciting elaborate responses when target users lack technology know-how [? ? ? ?]. Our probe extends an existing AI-driven application that diagnoses neonatal jaundice [?]. We implemented the probe (an AI-driven application) in Figma and instrumented it to "predict" neonatal jaundice instead of using actual AI to make predictions. CHWs used the probe to capture an image of a baby doll, receive a prediction, and view visual explanations for how the probe arrived at the prediction. The explanations were simplified versions of the two popular XAI methods used in human-centered XAI studies, LIME [?] and SHAP [?]. Using the probe helped CHWs to situate themselves as XAI users and critically reflect on the benefits and limitations of AI explanations in the context of an AI-driven tool that can be integrated into their everyday work.

Our findings show that the CHWs used their past experience with jaundice and other diagnostic devices (*e.g.,* thermometer, blood pressure monitors) to understand the functioning of the AI-driven probe. They struggled with the notion of uncertainty in the app's diagnosis and viewed it as a definite decision rather than a prediction, sometimes even doubting their own expertise in favor of the app's outcome. The SHAP and LIME explanations were difficult for them to understand because they perceived that the explanations described symptoms of jaundice instead of the importance of features on which the underlying AI relies to make predictions. The color-heavy nature of SHAP and LIME added to the confusion since they had strong preconceived notions of what different colors meant to convey. However, despite their confusion with understanding simplified XAI interfaces, CHWs were strongly in support of integrating explanations into the application and perceived several benefits of explanations, such as helping them learn new skills and

improve their medical knowledge, as well as helping them improve patient trust in AI tools and in themselves.

Given the high levels of AI overreliance and AI techno-determinism around such tools, we discuss the need to design new XAI methods that encourage users to think critically and skeptically about AI outputs and scaffolding structures that enable novice AI users to meaningfully and cautiously engage in cooperative work with AI-driven tools. Taken together, we make the following contributions to the field of CSCW:

(1) We examine how CHWs in rural India perceive and engage with AI explanations and show that the lack of interpretability can hamper effective cooperative work between low-skilled CHWs and AI-driven tools.
(2) By encouraging CHWs to engage with simplified versions of existing XAI methods, we gain in-depth knowledge about the discrete features of current XAI designs that shape their AI understanding.
(3) We present design recommendations for future XAI visualizations to improve the utility of AI explanations for novice AI users in high-stakes, non-Western contexts.

2 BACKGROUND AND RELATED WORK

We begin by providing background on community health workers (CHWs) and then describe the increasing integration of AI into their workflows. We then examine existing literature within the emerging area of research on explainable AI (XAI), detailing the XAI methods used within our study, describing research advances in human-centered XAI, and designing XAI methods for non-technical users.

2.1 AI and Community Health Work

Many developing countries in the Global South rely on the work of paraprofessional community health workers (CHWs) to provide last-mile healthcare in low-resource communities [? ? ? ?]. CHWs are usually women with a high school education who are recruited from local communities. After undergoing a few weeks of medical training, they then provide essential healthcare services like immunizations, family planning advice, and maternal neonatal care in hard-to-reach areas. They work in strongly patriarchal societies, where, as women, they have limited agency; hence, their labor is often unacknowledged, undervalued, and unregulated [?]. Yet, CHWs provide a critical link between the community and the public health service and are key to sustainable and resilient rural-urban health infrastructure [? ?]. They have been shown to positively impact healthcare outcomes, including reducing neonatal mortality rates [?] and positively changing behavior [?].

Many HCI and CSCW researchers have taken a broad interest in developing tools to support CHWs in their daily work. These tools cooperatively work with CHWs to help them collect data to aid in monitoring and evaluating their patients [? ? ?], track supplies for distribution [?], receive feedback on their work performance [? ? ?], and improve their knowledge and skills [? ? ?]. Increasingly, these digital tools have started to incorporate AI for disease diagnosis [?], scheduling visits [?], streamlining routine care [? ? ?], improving maternal healthcare outcomes [?], and analyzing rapid diagnostic tests [? ?].

A growing number of scholars have also examined the promise and potential of AI in addressing last-mile healthcare problems [? ? ? ?], which CHWs in low-resource areas often address. While AI holds promise in advancing the work of CHWs, many of these AI-enabled solutions are still in early deployment stages, and their potential impacts are not well understood.

The increasing integration of AI-driven tools into CHWs' workflows is concerning since prior work shows that CHWs perceive AI to be infallible and lend substantial amounts of trust in these systems [?]. For effective collaborative and cooperative work to happen between CHWs and AI-driven tools, it is important to examine CHWs' understanding of AI and make its inner workings understandable to them. Our research addresses this critical gap by analyzing CHWs' interactions with and understanding of AI explanations and the potential benefits and challenges of integrating explanations in AI-driven tools designed for CHWs. To motivate our research approach, we now look at prior work in explainable AI and the emergence of human-centered work within this domain.

2.2 Explainable AI

Explainable AI (XAI) consists of a set of methods that enable humans to understand the predictions made by machine learning models [? ?]. Typically, XAI methods may be applied in local, cohort, or global ways [? ?] to understand how the features within a model contribute to a single prediction, a set of predictions, or all predictions produced by the model, respectively. The focus of our paper is on local explainability, as individual predictions are considered to be most relevant for end users of ML models [?]. We now examine how researchers have begun approaching the concept of centering humans in the design, development, and deployment of XAI methods.

Human-Centered XAI. The field of human-centered design advocates for researchers to engage with their target users before developing and deploying novel technologies [?], however, these practices are not commonly engaged within the development of XAI methods [?]. Ehsan et al. [?] introduce the concept of "Human-centered Explainable AI" (HCXAI) as a method that centers human users when designing XAI tools. A review [?] of 85 HCXAI studies found that such studies measure five dimensions of XAI: trust, fairness, understanding, usability, and human-AI team performance. Our work focuses on understanding XAI by interacting with CHWs to examine how visual-based XAI methods impact their ability to interpret predictions.

Work in this space has focused on interviewing users and practitioners (UX designers, data scientists, researchers, etc.) to understand gaps in existing XAI tools [? ? ? ?], evaluating XAI methods with various stakeholders [? ? ? ? ? ? ?], and introducing participatory co-design of XAI systems [? ? ?]. Work critically examining the utility of XAI, especially in the scope of human-XAI collaboration, has employed user studies that demonstrate a misuse of XAI tools [?], a lack of assistance provided to users in model evaluation [?], the potential of XAI to mislead users into accepting incorrect decisions [? ? ? ?], and a limited ability for users to discriminate between correct and incorrect model predictions [? ? ?].

These risks of XAI could lead to drastic outcomes in high-stakes domains such as healthcare if, for example, physicians use explanations from incorrect predictions in their decision-making. To reduce overreliance on XAI, ?] argue for increasing people's cognitive motivation for engaging analytically with explanations and developing effective explanation techniques. ?] showed that overreliance can be reduced by lowering the effort needed to understand and verify explanations, improving the efficacy of human-XAI collaboration [?]. The work presented within HCXAI provides valuable insights that may generalize to wider populations, but AI and XAI are often used in very context-specific ways and with specific user populations. By relying primarily on crowdsourced participants who are primarily situated within the United States [? ? ? ? ?], there remains less understanding of how XAI could potentially impact users who make up the majority of the

world's population. Our study contributes to emerging literature in HCXAI by decentering Western perspectives on XAI to specifically focus on CHWs in rural India who are novice technology users and directly engaging with them through a field study.

XAI for Non-Technical Users. HCI researchers strongly advocate against one-size-fits-all approaches [**?** **?**] and emphasize the need to cross-validate principles and measures with different populations [**?**]. As XAI methods continue to develop, it is increasingly important to empirically analyze XAI techniques with a wide range of stakeholders ranging from AI practitioners to non-technical end users. Existing research in this space has focused on understanding how non-technical users form mental models when interacting with explanations [**?**], proposed methods to make models more explainable to non-technical users [**?**], and built frameworks to support AI and design practitioners in creating XAI prototypes that cater to non-technical end users [**?**]. In relation to our study, several researchers have designed and evaluated visual-based XAI approaches to improve understanding for non-technical users [**?** **?** **?**]. For example, Shen et al. [**?**] examined challenges the general public faces in understanding confusion matrices, a tool used to convey the performance of machine learning classifiers.

However, most of the existing work on XAI generally, and on non-technical users specifically, is focused on Western populations and contexts. Additionally, there is little work focused on XAI in the Global South, as evidenced in a review by Okolo et al. [**?**], which found that only a handful of studies have engaged with end users in the Global South to make AI explainable. Although a nascent but growing body of work has started to explore how concepts of AI fairness and explainability differ across Western and non-Western contexts [**?** **?**], little is known about what end users like CHWs, who have limited AI/digital literacy, need to know to cooperatively work with AI-driven tools in high-stakes settings. We contribute to this nascent domain by examining: **(1) How do CHWs engage with and perceive AI explanations in the context of AI-driven pediatric disease diagnosis?** and **(2) How do we design XAI interfaces that are more understandable to them?**

3 METHODOLOGY

To answer our research questions, we conducted a qualitative study with CHWs in rural India. To recruit CHWs, we partnered with a grassroots organization in Western Uttar Pradesh that runs several programs to strengthen community health systems in this region. A staff member from the organization contacted CHWs, explained the purpose of our study, and then scheduled interviews with the interested CHWs. All of our interactions with the organization and participating CHWs took place in person.

3.1 Field Procedure

HCI and CSCW researchers frequently use technology provocations [**?**], exploration artifacts [**?**], and cultural probes [**?**] to inspire users to think about new technologies and better understand their needs in real-world settings. Given CHWs' low familiarity with AI generally and XAI tools specifically, we designed a probe to let CHWs engage with an AI-driven tool to detect neonatal jaundice, in which explanations accompany AI's recommendations. This probe was inspired by an existing AI-driven application to detect neonatal jaundice, BiliCam [**?**]. We incorporated popular XAI methods into the instrumented probe so that CHWs could engage with a tangible artifact and feel encouraged to think critically and concretely about explanations instead of abstractly. We implemented the probe in Figma and *instrumented* it to emulate AI's predictions and explanations.

Observation and Interviews. We recruited 35 CHWs to interact with the probe and conducted semi-structured interviews with them. Three authors attended each session, with one leading the

Fig. 1. A participant interacting with the probe.

interview and the other two taking detailed notes and photos. The interviews took place in a closed room in community health centers to make it easier for CHWs to participate in the study in a natural environment. We first described the study to them and requested informed consent. After participants agreed, we then asked demographic questions to understand how long they had been working as a CHW, their smartphone usage, age, and education. We also asked participants about their knowledge of neonatal jaundice and how they diagnose it. We then described the functionality of the probe, explaining that it was designed to diagnose neonatal jaundice by analyzing an image of a child, and used an iPad Pro to show the Figma prototype. We opted to use an iPad instead of a smartphone for its larger form factor. To simulate how CHWs would use this probe in real-world situations, we provided life-size baby dolls and colored them yellow to simulate a real-life jaundiced child. Once a photo of a jaundiced doll was uploaded in the probe, it was *instrumented* to show: (1) jaundice prediction (including its severity level — mild, moderate, or severe), and (2) two XAI visualizations to explain the prediction. Our probe did not run an actual machine learning model to make predictions. Instead, the Figma prototype outputted predefined predictions and associated visualizations. Before each interview, we decided which prediction to show the next participant (e.g., moderate or severe jaundice).

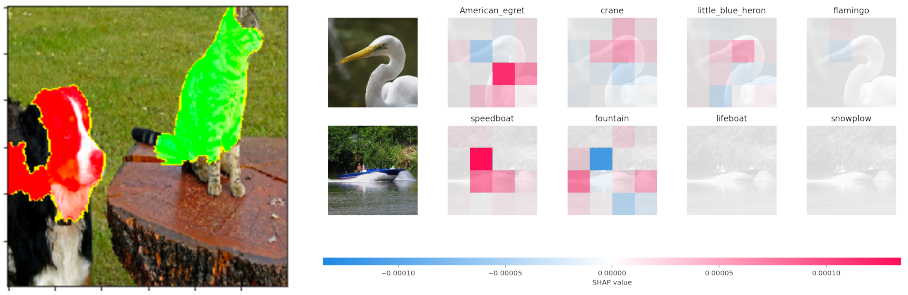
After we showed a demo to the participants, we asked them to interact with the probe and closely observed them (see Figure 1). Once participants came to the prediction screen (before moving on to the explanations), we asked them questions about their experience of interacting with the probe and their understanding of how the probe predicted jaundice. We then asked participants to "think-aloud" [**?**] as they interacted with the first XAI visualization. Following this, we asked them several questions to gauge their understanding of the XAI visualization, including how this explanation helped them understand the prediction and what they liked and disliked about the explanation. We then asked participants to interact with the second XAI visualization and repeated the process. At the end of their interaction with both XAI visualizations, we asked participants which explanation they liked better and why, and if they would find such explanations helpful if an app was developed to help them diagnose jaundice.

3.2 Probe Design

Representing Jaundice. We instrumented the probe to show moderate or severe levels of jaundice when a photo of a jaundiced doll is uploaded to it. Jaundice is a condition that causes yellowing in the skin due to elevated bilirubin levels in the blood. It is a progressive disease that affects the

Area of the body	Level	Range of serum bilirubin	
		μmol/L	mg/dL
Head and neck	1	68–133	4–8
Upper trunk (above umbilicus)	2	85–204	5–12
Lower trunk and thighs (below umbilicus)	3	136–272	8–16
Arms and lower legs	4	187–306	11–18
Palms and soles	5	≥306	≥18

Table 1. Kramer's Rule for visual assessment of neonatal jaundice [?].



(a) LIME (b) SHAP

Fig. 2. Default LIME and SHAP representations.

lower regions of the body as it increases in severity. In newborns, jaundice is classified on five levels based on Kramer's rule (see Table 1), which illustrates the relationship between the progression of jaundice and its accompanying bilirubin levels [?]. To simplify these levels in our representations, we depicted explanations of moderate and severe jaundice in a manner in which we focus only on the regions important to a specific severity level. We thus represented explanations of severe jaundice as yellowness in the hands and feet and moderate jaundice as yellowness in the shoulders, arms, torso, and legs.

Explaining AI Predictions: LIME and SHAP. LIME [?] and SHAP [?] are the two most popular methods used by AI researchers and developers to incorporate local explainability into their models, especially when designing for underserved communities in the Global South [?] and evaluating XAI in human-centered studies [?]. We thus chose to use LIME and SHAP to explain hypothetical predictions from the high-fidelity probe designed to diagnose neonatal jaundice. While LIME and SHAP are not designed for low-technical users, we used simplified versions of these methods because they provide visual/graphical interfaces for XAI, which are known to work better for low-literate populations [? ? ? ? ?]. Using LIME and SHAP in this manner allowed CHWs to think critically about AI explanations for disease diagnosis, which in turn allowed us to elicit elaborate responses through interaction with a tangible artifact.

The default representation of a LIME explanation uses green to highlight pixels of an image that positively contribute to the predicted class and uses red to highlight the regions that negatively contribute to the predicted class (see Figure 2a). The default SHAP explanation produces a visualization depicting the original input image (as a reference image) with the images of the predicted class and other classes following it (see Figure 2b). On these images are squares that highlight how specific features (pixels) of an image contribute to the respective prediction. There is a color

Gender	Female: 35, Male: 0
Age (years)	Min: 30, Max: 54, Mean: 42.8, Std: 6.6
CHW Experience (years)	Min: 4, Max: 17, Mean: 14.0, Std: 3.9
Technology Use	Computer: 0, Feature phone: 9, Smartphone: 26 (2 months – 10 years)
Education Level	Middle school: 13, High school: 15, Bachelors: 4, Masters: 3

Table 2. Demographic details of our participants.

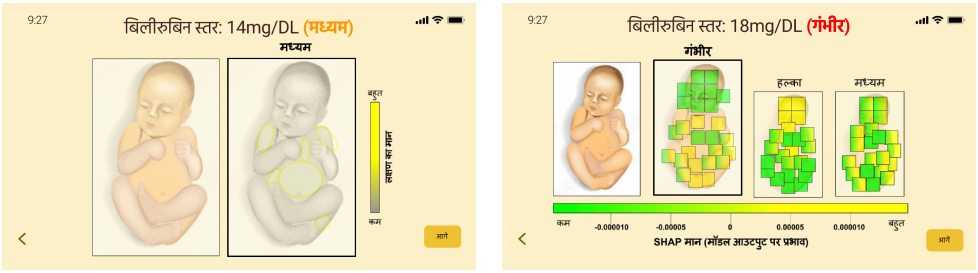
bar underneath these boxes with a gradient that ranges from blue (negative contribution) to red (positive contribution), indicating the respective SHAP value, which is computed on running the SHAP library. Since the visualizations designed for our study were not produced from a trained ML model, there is no reference to what "actual" SHAP values would look like for our specific use case of diagnosing neonatal jaundice. So, we used the same scale as provided in reference examples found in the SHAP documentation (see Figure 2b).

Simplifying LIME and SHAP for CHWs. The goal of our work was not to improve the interpretability of these methods but to inform design choices for building novel post-hoc XAI methods for novice AI users. To elicit specific feedback toward this goal, we simplified the original LIME and SHAP representations for CHWs. Before conducting the interviews, we organized a workshop with the staff of our partner organization to seek their feedback on the probe and the XAI visualizations.The staff expressed concerns regarding the use of red color in the default representations of LIME and SHAP. They were worried that the red color might confuse the participants since it is often used in public health messaging to imply "danger" in the medical sense. Since LIME traditionally uses red to highlight features with a negative contribution, non-jaundiced areas of the body would be red in the default representation, possibly indicating to our participants that such areas are contributing to the disease diagnosis. We thus changed the red (negative contribution) color to gray and the green (positive contribution) color to yellow, resulting in yellow for positive and gray for negative feature contribution in LIME (see Figure 3a). For SHAP, we changed the red color (positive contribution) to yellow and blue (negative contribution) color to green, resulting in yellow for positive and green for negative feature contribution (see Figure 3b). These color changes made our LIME and SHAP explanations slightly different from the original LIME and SHAP representations, but simpler for CHWs to understand. We also translated the probes into Hindi, which was done with the help of co-authors who are native Hindi speakers with rich experience working in community healthcare contexts.

While we had enough motivation to change the colors and language of LIME and SHAP visualizations, there was not enough relevant literature to justify any other simplifications. Hence, throughout our study, we iteratively updated the XAI visualizations in response to the feedback we received from CHWs to make the explanations more understandable to them. Examples of such changes included: resizing reference images, using different shapes (squares and circles) instead of colors to indicate feature importance, and creating more descriptive text labels. Examples of these changes are shown in Figure 5 and Figure 6. We describe the rationale for these updates in further detail and their impact on CHWs' understanding in Section 4.

3.3 Participants

The CHW participants in our study lived and worked in rural Uttar Pradesh, one of the poorest states in northwestern India. They performed their work by traveling door-to-door, visiting patients in their homes, and providing them with family planning advice, maternal and neonatal care, and essential health services based on approved health protocols. All of our participants were women, which is standard for practicing CHWs in India. They ranged in age from 30 to 54 years



(a) LIME (b) SHAP

Fig. 3. Screens of the explanations used in our probes. At the top, the screens state the "bilirubin level" and the jaundice prediction: "mild" in (a) and "severe" in (b). In (a), yellow is used to highlight the bodily regions of the baby that positively contribute to the severity of jaundice. The colorbar on the right serves as a legend for the "feature importance" from "low" to "high." In this case, since the shoulders, arms, torso, and legs are highlighted in yellow, the predicted class is "moderate." In (b), yellow boxes indicate pixels with higher SHAP values and green boxes indicate pixels with lower SHAP values, as explained by the colorbar labeled "SHAP value (impact on model output)." Higher SHAP values indicate a higher feature contribution to a respective class prediction. The three annotated images are labeled "severe", "mild", and "moderate" in decreasing order of the model's confidence. In this case, since the yellow boxes are primarily situated over the hands and feet, the predicted class is "severe", written in bold above the largest image. Since an ML model was not deployed in this study to output SHAP values, the numbers used on our plot are lifted from the default SHAP representation in Figure 2b.

old and had an average of 14 years of experience working as a CHW. All of them owned or had access to a cellphone, with the vast majority (75%) using smartphones. Most of our participants had limited digital literacy and low levels of AI knowledge and exposure. None of them used a laptop or computer. The majority of CHWs (80%) had at most a high-school education, with 13 CHWs completing middle school, 15 completing high school, 4 with a bachelor's degree, and 3 with a master's degree. We summarize the participant demographic information in Table 2.

3.4 Data Collection and Analysis

Throughout our fieldwork, we collected 15 hours of audio recordings and 57 pages of detailed notes. Audio recordings were translated into English and transcribed. We then used inductive thematic analysis [**?**] that allows key themes to emerge from the raw data through repeated examination and comparison. Two authors led the qualitative coding process. To begin, they each coded three interviews separately and met to reconcile conflicts by merging similar codes and organizing the remaining codes into a streamlined codebook. They continued to code individually and came together after each interview until we had reached intercoder reliability and codebook stabilization. Following this, they separately coded the remaining transcripts. Throughout the analysis, the research team held multiple discussions to refine the codes iteratively and used peer debriefing to reconcile disagreements. After conducting multiple passes, we ended up with 121 codes grouped into five themes: General XAI Interpretations, Probe Interpretations, XAI Preferences and Suggestions, SHAP, and LIME. For example, the "General XAI Interpretations" theme categorized CHW interpretations of features that were consistent across all explanations, like bar graphs, colorbars, and images. Examples of such codes include: "XAI Interpretation: bar graph", "XAI Interpretation: gradients," and "XAI Interpretation: reference image."

3.5 Ethical Considerations

We gained IRB approval for all study procedures. Given the COVID-19 pandemic, we were vigilant in protecting the safety of the study participants. We also used face masks and hand sanitizer and sanitized the study implements (iPad and baby dolls) regularly. We were also flexible in scheduling the interviews since CHWs had to juggle their schedule to provide patient care. Participants were remunerated in kind by our partner organization for their involvement in the study.

Positionality Our mixed-gender team comprises researchers who are from countries in the Global South and have rich experience in conducting fieldwork with underserved communities in India and other low-income regions. Two of us have over a decade of experience studying CHWs in South Asia and Africa, and one of us has worked with our partner organization for several years. Despite this shared background and experience working with them, we are not low-income and have not lived in rural regions for extended periods. Thus, our education, gender, socioeconomic status, and urbanity placed us in a position of power with participants who were low-income women living and working in strongly patriarchal systems in rural regions. To facilitate a comfortable environment, our mixed-gender team conducted interviews in health centers (workplaces for CHWs) in the presence of a woman field staff of our partner organization. This approach, together with our past engagements with CHWs in rural communities, helped elicit in-depth responses from the participants. We approached this work from an emancipatory action research mindset to understand how socially and culturally diverse CHWs in the Global South engage with XAI methods and how this knowledge might inform design recommendations to improve the utility of XAI methods for novice AI users in non-Western contexts.

4 FINDINGS

All CHWs struggled to understand AI explanations. While this was expected, through a deep engagement with the probe, they provided elaborate responses on how they perceive the benefits and pitfalls of AI explanations and how different features in XAI visualizations impact their understanding.

4.1 How CHWs Interpreted AI

At the beginning of the interviews, we asked participants if they were familiar with AI. Irrespective of their experience with technology, most CHWs had not heard of AI and did not know what it is. After we showed the participants the app (probe), we asked them to describe how they think the app diagnoses jaundice in newborns. While almost all CHWs understood what the app does (e.g., help diagnose jaundice in babies), they struggled to pinpoint how the app works. In particular, they were unable to *explain* how the app arrived at the prediction, often being clueless and comparing the process to *"magic."* Only a few CHWs were able to guess that the app must be looking at the skin color to detect yellowness, particularly since they had no exposure to end-user applications that use computer vision.

Since the CHWs could not understand how the app arrived at a prediction, they relied on their knowledge of jaundice and neonatal care to interpret the prediction and accompanying explanations. CHWs would often look at the doll on the table and use their knowledge of expected "symptoms" to make sense of the prediction: *"The baby's feet and stomach looks yellow, that's why the app has encircled that region"* (P35). They often used their perception of how the baby (doll) is "feeling" as a key factor in their individual decision-making, stating things such as *"The doll looks cranky, that's why the app has detected jaundice"* (P18), something which is very difficult for AI to

do in practice [**? ?**]. In a way, the CHWs were ascribing superior capabilities to the AI-driven app than it actually possessed.

Comparison to Diagnostic Devices. CHWs' prior experience with medical diagnostic devices shaped their AI-related mental models. Throughout the interviews, CHWs often relied on their experiences working with diagnostic devices to understand how the app might be detecting jaundice. One participant compared the probe to blood pressure monitors and thermometers and stated, *"How can I tell how these machines work? They just do."* (P02). Similarly, P04 stated, *"The same way a thermometer checks for fever, this app is checking for jaundice."* CHWs stated how in most of the medical equipment and diagnostic tools they use in their daily work, no explanations are provided in the output, and they also do not know how such devices work under the hood. For them, a thermometer simply provides a temperature, a scale provides a weight, and a COVID-testing system gives a diagnosis. Additionally, CHWs stated that detailed reports from medical tests are commonly interpreted by a trained doctor, not by them. Hence, the concept of needing to understand exactly how the probe comes to a specific jaundice prediction was quite discomforting to our participants. However, some CHWs conceded that they need to know *enough* to operate these devices safely and understand the output to a degree where they could explain it to the community members. "What" these diagnostic tools do was more important to them than "how" they do it. These findings show that when CHWs interact with AI-enabled tools designed to augment their work, they are prone to rely on their prior experience operating medical devices and providing patient care to discern how such tools operate.

Discomfort with Uncertainty in a Machine's Output. We also observed that the CHWs often treated the output from the probe as an absolute decision instead of an estimation or prediction. They assigned the same level of trust to the AI app as they would to any diagnostic system like a thermometer, blood pressure monitor, or weighing scale, believing that the app could never be wrong. For example, P08 emphasized, *"a machine can never be wrong"* and P15 felt that *"since the app is a computerized system, it is natural that it would give the right result."* Moreover, when the explanations showed them things that went against their intuition, instead of questioning the prediction and the underlying AI, they started doubting their own knowledge and understanding of jaundice. Past work has also uncovered overreliance on AI, *e.g.*, accepting incorrect decisions without verifying whether the AI is correct [**?**]. However, this problem becomes more severe when low-skilled CHWs use nebulous AI systems in high-stakes settings where results must be interpreted with caution.

4.2 How CHWs Perceived XAI Explanations

In addition to understanding participants' perceptions of the diagnostic capabilities of the app, we also explored their reactions to the provided explanations. Most CHWs struggled to understand the explanations of how the underlying AI arrived at decisions. Although a few participants stated to understand the explanations presented to them, the understanding they voiced varied greatly from the intended meaning, and they kept changing their interpretations every time we asked for their understanding. For example, P17 said the two images in LIME were of the same baby but later said that the baby in the second image had more severe jaundice than the baby in the first image, contradicting her claim that they were photos of the same baby. Other CHWs thought that the SHAP explanation showed a story of a baby transitioning from having no jaundice to severe jaundice. While these interpretations made sense to the CHWs, they were not what the XAI methods intended to convey (they intended to show the original image for reference and the annotated image for explainability, with no storyline among the images). Only a few CHWs could partially understand the explanations with the help of higher-level features such as colors, images,

or shapes. For example, P07 understood that the SHAP explanation was showing varying severities of jaundice: *"In the severe image, the entire body has yellowness. In the mild one, it's only the face. In the moderate case, the entire body has yellowness, but it is light yellow"*. Almost all CHWs explicitly stated their confusion, often asking the interviewer to help them understand the explanations. We also found that most CHWs already placed high trust in the AI prediction, and the presence of explanations further reinforced their trust in the app. These outcomes are concerning because existing research has shown the possibility of explanations to reinforce incorrect predictions [**?**], which in the case of healthcare could have severe consequences.

Symptoms vs. Feature Importance. The CHWs construed the XAI designs to depict the *symptoms* of jaundice in the baby rather than depicting important contributors to the app's prediction. To them, the visualizations showed parts of the baby that have jaundice, not parts of the baby that led to the prediction. For example, P05 noticed the outlines on the LIME representation and said the line represents parts of the baby infected with jaundice. In SHAP, P07 interpreted the yellow boxes as body parts infected with jaundice and the green boxes as healthy body parts. In reality, while some body parts may appear more yellow than others, jaundice does not affect different body parts with different severity. The entire body is said to have jaundice, not isolated body parts like the hands or feet. All CHWs displayed this correct understanding in isolation, but it got overshadowed every time we discussed the XAI visualization, saying that some parts of the body had "more" jaundice than others. This suggests that XAI visualizations caused confusion between feature importance versus symptoms of jaundice. For example, P35 got confused when we asked her why the baby's foot is not yellow. She replied, *"There is less jaundice in the foot, more in the rest of the body."* Like most other CHWs, she alluded to yellowness in the body, which is a mere symptom of jaundice. In reality, the lack of yellowness in the foot indicated to the model that the baby might have moderate jaundice and not severe jaundice since yellowness in the foot is a symptom of severe jaundice as per Kramer's rule (see Table 1).

4.3 Why CHWs Still Wanted Explanations

Although CHWs struggled to understand the explanations, they still expressed a strong desire to have explanations. They mentioned various advantages, such as explanations improving their comprehension of the diagnosis, enhancing community trust in their work, and helping them explain how the app works.

Better Understanding of Diagnosis. When we asked CHWs why they preferred to have explanations, they mentioned that the explanations would allow them to understand how the app arrived at a diagnosis, especially once the explanations are made easy to understand or if CHWs are trained to interpret them. P09 emphasized the utility of the explanations in making her understand the diagnosis but warned against their interpretability:

> *"See, they both [LIME and SHAP] are very confusing. I think you should take it out of the app. Keep it only if we can understand. If it is there and is easy to understand, then it will have value. We will have more information, and we will not forget things."*

This shows that explanations are useful to CHWs only if they can be interpreted. CHWs received only a few weeks of medical training and thus saw a huge potential in explanations to help upskill them. Many CHWs believed that an AI-driven tool in which explanations accompany AI's recommendations would serve as a tutor, teaching them new skills and providing them with hands-on training to detect jaundice.

Improved Community Trust. Early work in HCI and ICTD [**? ?**] has shown the importance of mobile phones in offering a sense of "legitimacy" to the work of CHWs, and more recent work [**?**

Fig. 4. An example of paper-based visual aids that CHWs used to improve community trust in them.

] has demonstrated how AI could also render more authority to CHWs but potentially decrease patient trust if AI tools produce incorrect predictions. It was of great importance to our participants to be able to relay detailed information about a diagnosis to the broader community. Since no such applications are currently used in the community, CHWs felt that the explanations would demonstrate that the app does not arbitrarily produce decisions and increase the community's trust in the app and in them. P03 highlighted how the explanations would prove that the app did not arrive at a decision randomly and engender trust in them:

> *"People will easily understand how the decision is taken if the app is there...and that they need to see the doctor ASAP."* (P03).

P04 and P05 emphasized that their patients trust them more when they use visual aids. They gave the example of paper-based visual aids (see Figure 4) they carry to educate pregnant women about when to seek urgent care. When we asked if CHWs would prefer similar paper-based visual aids for jaundice, P04 and P05 preferred the AI-driven app over paper-based visual aids, stating that the app could diagnose the severity of diseases, which the paper-based aids could not. P06 shared:

> *"The app is useful because it tells how much the baby has jaundice. I can't do that. The public would also trust the app more...If we use the app, we will also learn how the app diagnoses jaundice, which will also help us improve and share the information with community members."* (P06)

CHWs mentioned that the explanations would also allow parents to take a diagnosis more seriously. As P08 put it, *"they will understand why their baby has jaundice and what are the associated symptoms."* While this was perceived to be a beneficial outcome of including explanations, some participants debated the right amount of details and the language to be included in the explanations to prevent their patients from experiencing distress. P12 argued, *"explaining too much would cause the villagers to worry or panic."*

4.4 How Color Impacted CHWs' XAI Understanding

Color was prominently used in both LIME and SHAP explanations. In this section, we detail the usage of color within LIME and SHAP and how it impacted the CHWs' understanding of XAI visualizations.

Colorbars. Colorbars were used in the XAI visualizations to indicate the contribution of an underlying feature to the model's decision-making. In LIME, the colorbar was situated vertically with a color gradient indicating low feature importance at the bottom and high at the top (see Figure 3a). In SHAP, the colorbar was situated horizontally with a color gradient indicating low feature importance at the left and high at the right (see Figure 3b). Only one out of 35 CHWs was able to use the colorbar effectively to form an understanding of the XAI visualizations. The rest struggled to interpret them. While some participants (e.g., P07, P17, and P18) noticed that there was more yellowness at the top and less at the bottom of the LIME colorbar, they were unable to relate this information to form an understanding of the LIME visualization. P17 perceived this difference in color to mean that the baby would still have some chance of surviving unless the colorbar became entirely gray, while P04 incorrectly interpreted the colorbar as a scale of severity of the disease:

> *"There is more yellowness at the top and less at the bottom... Three-fourths of this bar is yellow, and the remaining is gray... I think this means that out of four parts, the kid has jaundice in three parts."* (P04)

While CHWs noticed the text accompanying the colorbars, it did not improve their understanding. For example, P17 incorrectly interpreted the meaning of "less" to mean that there is a lack of blood and "more" to mean that there is more yellowness in those areas.

Mental Models of Color. CHWs often relied on their prior experience with colors in visual aids in the domain of Public Health to interpret what the colors in the XAI visualizations might mean. They had strong color associations, believing that *"green means safe, yellow means warning, red means danger and gray/blue means lack of blood."* These color associations largely influenced how they interpreted the visualizations depicting a jaundiced baby.

While interpreting the LIME explanation (Figure 3a), P06 felt the baby was in moderate danger because the body had patches of yellow color. When we asked why the face of the baby was gray, P06 felt that this was because the baby had a darker skin color. This interpretation was incorrect because the baby's true skin color was only visible in the reference image, and the LIME image was overlaid by the colors of the visualization (yellow or gray, in this case).

Some participants also noticed that the outline of shapes was darker than the fill color of the shapes. For example, the fill color was light yellow, and the outline color was a darker yellow to mark it off from the rest of the body. P20 noticed this subtle difference and interpreted that, *"since the outlines are darker, jaundice is more severe in the outer areas, and it is trying to spread more to the rest of the body."* This shows that CHWs tried to interpret even subtle details of the visualization. They placed immense importance on the explanation, expecting every minor detail to mean something. Hence, the design complexity lies in not just conveying what the explanation means but also avoiding what it doesn't.

To further simplify the visualizations, we tried exploring a color scheme that avoided these strong color associations. Through discussions with CHWs and the staff of the partner organization, we designed a new LIME visualization (see Figure 5a) that used a pink color scheme: dark pink to represent a positive impact on the prediction and light pink to represent a negative impact on the prediction. Even though pink seems close to red, it did not evoke a feeling of 'danger.' Instead, participants assumed that the dark pink color suggested a more severe level of jaundice in the highlighted body parts. While this was a partially accurate assumption, it was not something

(a) LIME (b) SHAP

Fig. 5. Additional re-designed LIME and SHAP explanations. In (a), some body parts are encircled in a dark pink color, and the rest of the body is shaded in light pink. On the right is a descriptive legend saying, "Which body parts contributed to the prediction of moderate jaundice?: [Dark Pink]: More contribution, [Light Pink]: Less contribution." In (b), the three images show the three severity classes labeled "moderate", "mild", and "severe" jaundice. The current predicted class is in bold ("moderate") and of the largest size, with each following image smaller in size, depicting decreasing confidence in those classes. Green and yellow squares are on each image. The legend on the right contains exactly the same text as in (a), with green squares for less contribution and yellow ones for more.

(a) LIME (b) SHAP

Fig. 6. Re-designed LIME and SHAP explanations without colors. In (a), the circular outlines lead to a descriptive text saying, "These body parts contributed more to the prediction." The text box on the bottom says, "Rest of the body contributed less to the prediction." In (b), the three images show the three severity classes labeled "moderate", "mild", and "severe." The current predicted class is in bold ("moderate"). Above each image is a confidence label, "most", "a little less", and "least" confidence. On each image are shapes: squares representing less contribution to the prediction and circles representing higher contribution, as explained by the descriptive legend on the right.

that the visualization meant to convey. Given that the CHWs had weaker existing mental models of pink being used in the context of disease diagnosis in their daily work, their assumptions were more measured.

Prototypes Without Color. Since even colors like pink did not elicit a correct understanding, we later simplified the probe not to use any colors.

In SHAP, we replaced colored boxes with shapes: circles for higher feature importance and squares for lower. To accommodate this change, we removed the colorbar and added a discrete legend with a descriptive title (see Figure 6b). Our main aim was to design XAI interpretations that were not affected by CHWs' mental models of color. We did not want them to interpret one color as "more jaundice" and another color as "less jaundice" in the body parts, but instead as

"more" or "less" contribution of the body part in the AI's prediction (i.e., *feature importance*). Upon using these shapes, CHWs started to move away from their jaundice/no-jaundice interpretation. Since they did not have colors to base their explanations upon, the participants started to use their understanding of jaundice to interpret the visualization. For example, P29 pointed to the circles and said that in her experience, the encircled areas most commonly help with diagnosing jaundice. Further, since the shape did not have a color fill, CHWs could now notice that the baby's encircled body parts looked yellow and use this information to guess the severity of the disease. With some hand-holding, CHWs that were shown this probe were able to get an initial understanding of the correct interpretation of SHAP.

In LIME, we also designed probes that tried to depict the same information but without using color. As seen in Figure 6a, in this version, we outlined the parts of the body with less and more contribution and labeled them with text descriptions. While there was some inertia in the CHWs to read large blobs of text, once they did, some of them started to form an initial understanding of the explanation. Devoid of color, this version made it easier for some CHWs (P25, P33, P35) to move towards a more correct understanding of what LIME intended to explain. P33 said:

> *"The stomach, feet, chest, and shoulders [encircled body parts] help in identifying moderate jaundice. These body parts are playing a larger role in telling that the baby has jaundice. The rest of the body parts are informing the computer less about jaundice."* (P33)

However, there was still confusion about the specifics. While the size of a respective circle was not indicative of feature importance, one participant mentioned, *"The stomach has the most contribution because a bigger circle is drawn over it."* (P31). In reality, the stomach only had a bigger circle because it was a bigger region.

These findings show that removing color from the visualizations was helpful for some CHWs. It prevented them from relying on their existing mental models of color associations, thereby improving their understanding of how the probe estimated jaundice severity. However, only a few CHWs benefited from this change. Most participants continued to view the encircled body parts as symptoms of jaundice rather than the explanations of the prediction.

4.5 How Graphical Elements Impacted CHWs' XAI Understanding

Multiple Images on the Screen. The LIME visualization in the app displayed two images of the baby (see Figure 3a): the input image and an annotated image describing feature importance. The SHAP visualization displayed four images of the baby (see Figure 3b): the input image and an annotated image for each severity class (mild, moderate, and severe). The input image is unlabelled and serves as a reference image, and the three annotated images represent the model's evidence for the three severity classes in decreasing order of its confidence. Multiple images of the baby in the same visualization caused a lot of confusion for our participants.

For example, CHWs were often confused by the input image. We simplified our subsequent prototypes by writing the literal text "The image you entered" in Hindi on top of the reference image, but even then, most CHWs were unable to understand that this image was only for reference. Instead, they provided varying interpretations, such as the image sequence demonstrating the progression of jaundice (P03, P06), representing multiple babies with different levels of jaundice (P04, P05), or the baby in the reference image being *"less"* sick than the ones in other images (P06, P07, P23). P06 stated:

> *"The baby in the reference image looks weak and inactive. But this baby has no jaundice because the hands and legs are relatively less yellow."* (P06)

In the SHAP visualization, the CHWs interpreted the labels for the other three images literally and assumed that the baby labeled 'mild' had mild jaundice and the one labeled 'severe' had severe jaundice. However, these severity labels meant to represent explanations for other classes had the model actually predicted those severity classes. For example, P04 and P05 felt that the baby had a mild infection initially but was at risk of getting a moderate or severe infection if left untreated. Similarly, P28 assumed that the images represented a chronological progression of jaundice: first, the baby was sick (moderate label), it was given medicine to get better (mild label), but then relapsed into severe jaundice (severe label).

Further, some CHWs tried to correlate the severity label to the *number* of green and yellow boxes in the respective image. In the case of the 'severe' jaundice prediction, they assumed that the baby labeled severe had severe jaundice because it had the most yellow boxes on it. This interpretation was incorrect because the underlying model relied on specific bodily regions (in alignment with the Kramer's Rule chart in Table 1) to provide a severity prediction, not on the number of yellow boxes. For example, a "severe" prediction would primarily have yellow boxes covering the palms and soles of the feet compared to a "moderate" prediction with yellow boxes covering the shoulders, arms, torso, and legs. In the case of the moderate jaundice prediction, the image labeled moderate had more yellow boxes than the image labeled severe (as more parts of the body are used by the model in the moderate prediction). Some CHWs found this confusing and pointed it out to us, either suggesting that the images are labeled wrong or that their understanding is flawed. Other CHWs did not notice this disparity, and when we pointed it out to them, they were left confused about their understanding.

Number Line. Only a few CHWs noticed the number line below the images in the SHAP visualization. For example, P02 observed the numbers on the line increasing from left to right but assumed that those numbers represent the severity of jaundice. Instead, the numbers indicated a quantitative (negative, positive, or neutral) feature contribution of the green and yellow boxes. P02 also interpreted the placement of the images as being labeled in increasing order of severity. However, the images were not labeled in increasing order of severity but in decreasing order of the model's confidence in the predicted severity (e.g., 'severe' was displayed before 'mild' when the prediction was 'severe'). To justify this inconsistency, some of the CHWs stated that the *"images were displayed in an incorrect order."*

4.6 How Textual Elements Impacted CHWs' XAI Understanding

Both LIME and SHAP visualizations had text placed at the top of the screen to indicate the app's prediction. CHWs almost always read this text but did not pay attention to other textual elements. Instead, they primarily focused on the graphical elements on the screen. For example, in our final LIME prototype, we removed all colors and the reference image and had just one image with an explanatory text (see Figure 6a). Even then, CHWs did not read the text without being asked to and only focused on the image. We often asked CHWs to read the text accompanying colorbars, legends, etc. Even though over 60% our participants had at least a high school education, almost all of them were unable to understand the mathematical textual elements. They did not notice the minus sign and the decimals used in the number line in SHAP, for example, many CHWs read -0.000010 as just 10.

Confidence Values in SHAP. As described earlier, the images representing different severity classes were presented in decreasing order of AI's confidence, which greatly confused our participants. In the simplified versions of the probe, we added explicit confidence labels to the images: as percentages in one probe (e.g., '95% confidence') and as descriptive text in another probe (e.g.,

'least confidence'), see Figure 6(b). However, these labels also did not improve the CHWs' understanding of the confidence intervals. For example, P28 ignored the % sign and interpreted "95" as body temperature since body temperature is usually measured in Fahrenheit in India and 95° is slightly below normal. However, she soon got confused as 3° (for '3% confidence') and 2° (for '2% confidence') did not make sense as body temperatures. P24 tried to interpret the confidence percentages slightly differently but got confused along the way:

> *"These numbers show decreasing symptoms of jaundice: 95, then 3, then 2. But why is the moderate image listed under 95 and the severe under 3? I don't understand."* (P24)

Our next attempt at simplification was to change the sizes of the three images such that the decreasing size of an image represents decreasing confidence. For example, the most confident prediction shown is the largest image, and the least confident prediction is the smallest image (see Figure 5(b). Most CHWs did not notice the different sizes of the images. When we asked them, the varying sizes did not help them interpret the confidence values. Many CHWs (e.g., P20, P21, P22, P24, P25) interpreted different sizes to represent children of different ages and assumed that the probe told them that jaundice could happen to kids of any age. P28 was left utterly confused:

> *"If the first image was the smallest, then I would understand that the baby is growing in size with age, but here, the first baby is the biggest... Oh, maybe the baby is getting drier due to the illness [meaning that its weight is reducing]?"*

By observing their experience with LIME and SHAP, we learned that the meaning of confidence itself was unclear to our participants in this context. When asked to guess the meaning, P23 said that *"these labels show what people in the field should trust more,"* which is true as an outcome of the confidence values, but not their explicit meaning in the XAI visualization. These interpretations also shaped our choice to call the probe's output a *prediction* instead of a *decision* to enforce the mental model that AI is merely *predicting* a jaundice severity and not providing a firm diagnosis. Our findings suggest that such actions might be necessary to help CHWs retain their agency when acting on predictions produced by ML models in high-stakes situations.

5 DISCUSSION

In this section, we begin with a discussion on understanding how to build XAI that meets users' needs. We then introduce the concept of using "scaffolding structures" to support users like CHWs in understanding intricate explanations and explaining AI to low-literate users by leveraging visualization-based XAI methods. We end by providing design recommendations that can inform the development and evaluation of XAI methods targeted toward novice AI users.

5.1 How can we build XAI that meets user needs?

Many CHWs believed that XAI would help improve their understanding of disease diagnosis and also guide their interactions with patients. As a result, they expressed a strong need for explanations to be interpretable by them. Indeed, many of the benefits touted by XAI are currently reserved for the developers of these models [**? ? ? ? ?**]. Additionally, these methods can provide conflicting explanations [**?**], encode trust in incorrect decisions [**? ?**], and increase model complexity [**?**]. To work towards improving the utility of XAI, AI practitioners and researchers will need to be aware of these constraints to build XAI that meets user needs.

Increasing User Agency. User agency is important in ensuring that humans working collaboratively with AI tools can be empowered to challenge AI decisions and decide whether to use predictions from these tools in their decision-making processes [**?**]. With prior work demonstrating the impact of XAI to lead to human overreliance on AI [**? ? ?**] and our work showing that

CHWs have the potential to defer to explanations, these findings potentially highlight the ability for reduced human agency in human-XAI interactions. As users increasingly become exposed to AI and are expected to make decisions with these tools, it is critical to understand ways to maintain and even potentially strengthen user agency to combat the negative effects of XAI. One way to increase end user agency when interacting with AI systems is to examine how XAI methods can help users think more *critically* about predictions produced by AI systems and the impacts of using these predictions in decision-making. Past work has reported over-reliance on AI in rural India [**?**] and the presence of "AI authority" [**?**] where users are more likely to accept AI decisions without fully understanding the capabilities of AI. Current explanation methods do not give end users any level of agency to rebut incorrect predictions, something that researchers and legal frameworks such as The EU General Data Protection Regulation (GDPR) frame as a right for users of AI systems [**? ?**]. New XAI methods that give workers a reason to question, engage, and think critically about AI need to be developed. A novel XAI method could be designed to understand how CHWs typically handle medical uncertainty in the field and mirror that process. For example, when CHWs receive inconclusive readings from rapid diagnostic tests, what steps do they take to mitigate these issues? Do they administer multiple tests, use their own knowledge of patient symptoms to estimate a diagnosis, or do they refer the patient to a clinic? By examining these questions, researchers can then understand if similar steps should be taken with X/AI systems and how end users like CHWs perceive the prospects of doing so. This "uncertainty procedure" could then be combined with contrastive explanations [**?**] to provide real-world examples (based on CHWs' prior experiences) as to why a particular diagnosis was given instead of another. Such methods could help prevent users from forming false beliefs about explained ML predictions [**?**] and provide these users with the agency to actively challenge these systems.

Accessible X/AI Terminology. Research continually emphasizes the need for XAI to center stakeholders while acknowledging their distinct needs [**? ? ? ? ? ? ?**], however current approaches to XAI methods drastically fail to do so for populations and contexts in the Global South. For example, as most AI/ML vocabulary is presented in English, considerations regarding the choice of language also have to be explored. The language barriers that existed when we translated the textual aspects of LIME and SHAP (colorbar labels, image labels) from English to Hindi may also hold the same for other languages not commonly represented in ML. For example, भविष्यवाणी ("bhavishyavaanee"), the closest word to "predicted" in Hindi, also translates literally to "prophecy." One CHW (P10) in our study interpreted this to mean that the app probe provides an uncertain diagnosis of jaundice, and the *actual* diagnosis would be given by a doctor. Researchers could potentially fill this void by collaborating across regions to create a working dictionary of X/AI jargon with simple definitions and translations into local languages and dialects. Prior work has created a dictionary of AI terms but such dictionaries date to the 1980s [**?**], 1990s [**? ?**], and only focus on translations to Western languages like English, German, French, and Italian [**?**]. Such an X/AI dictionary would expand upon the efforts by Skočaj et al.[**?**] that built a dictionary of AI terms in Slovenian and a series of YouTube videos by Wuraola Oyewusi [**?**] introducing AI concepts in Yoruba, a language widely spoken in Nigeria. The establishment of such resources would enable future developers of XAI toolkits and libraries to incorporate these translated terms, potentially serving a wider range of linguistically diverse users in non-Western contexts.

Understanding the Explanation Needs of End Users. We find that existing explainability methods are inept for what CHWs need, mainly because these techniques are primarily designed for ML developers and expected to generalize across a broad range of use cases. CHWs' need for XAI differs from developers' need for AI, which would also be different from the needs of

regulators governing how AI is used in healthcare. The major frustrations relayed by our participants mainly focused on their lack of understanding regarding basic AI tenets and mathematical concepts. Our study highlights that CHWs tended to avoid reading text and could not interpret numbers, decimals, and negative signs. In future implementations of XAI that are designed for use by CHWs, these features will need to be eliminated. As AI-driven tools are integrated into the daily workflows of users like CHWs, who lack AI know-how and have limited domain expertise in advanced medical diagnoses, new explainability methods that simplify or remove complex math and statistical terms are only the first step.

We also need to carefully understand what users like CHWs who have low AI literacy need to know in order to safely operate these devices in high-stakes settings to minimize the potential risks and harms. For example, we found that CHWs wanted AI to tell them the symptoms that impacted the prediction of jaundice severity instead of highlighting body parts that the AI model used to make a prediction. This would allow them to better understand the model's reasoning behind the prediction and provide them with more information to relay to patients. With this in mind, it is also important to understand how to explain AI to users who could be asked to explain the inner workings of AI to others, as seen in our findings and in work by Tonekaboni et al. [**?**] where clinicians view explanations as a potential method to justify medical decision-making guided by AI systems.

5.2 Scaffolding Explanations

Given low levels of AI knowledge and exposure among CHWs, it cannot be expected that CHWs will develop capabilities to understand intricate explanations themselves. It is critical to provide scaffolding structures that support them in understanding how AI operates "under the hood" and what realistic expectations they should have when working with AI-driven tools. Many of our participants expressed that training on understanding explanations would improve their utility for real-world use. For example, when users are exposed to apps or computer programs for the first time, they receive explicit training in the form of short-form tutorials or guided interactions. We believe that similar training should be incorporated to guide users when interacting with AI tools and XAI interfaces. However, while training users on how to interpret XAI may help address existing challenges with AI in the short term, such training should complement but not replace efforts towards developing more user-friendly XAI.

Despite their limited domain expertise, our study shows that CHWs have critical perspectives of existing XAI methods that can substantially impact the design of future explainability methods. Sambasivan et al. [**?**] specifically note the importance of domain experts like CHWs in high-stakes AI applications, advocating for their inclusion in the end-to-end development cycles of AI projects. To understand the type of support and training CHWs may need to operate novel explanations, researchers should intentionally work with CHWs to identify their needs. Compared with traditional approaches taken in developing and deploying ML systems, the HCI community strongly recommends using communities' existing methods and frameworks and building on them to expand to new areas rather than using one-size-fits-all approaches [**? ? ? ? ?**]. In this study, our participants found it particularly hard to understand the concept of uncertainty as it directly conflicted with their existing mental models of the devices they use daily in their work. The concept of AI being a *prediction* rather than a *measurement*, such as a weight or a temperature, skewed their understanding of explanations. To bridge this fundamental difference in thinking, future work could focus on developing support programs that build upon CHWs' existing knowledge in ways that make sense for the community, for instance, by using contextually relevant analogies and examples or by turning to their prior experiences with medical uncertainty where equipment failure and complex patient symptoms can lead to indecision [**? ? ?**].

5.3 Explaining AI through Visualizations

Within our work incorporating visualization-based XAI methods, we found that participants' mental models of colors often impacted their interpretation of how the app diagnosed jaundice. Other graphical and textual features like boxes, reference images, labels, and legends were also widely misinterpreted. Further, our efforts to remove color improved understanding and allowed CHWs to rely more on their domain knowledge than colors to interpret the explanations. Such slight changes show promise in aiding the understanding of XAI for diseases similar to jaundice, which are impacted by discrete features such as color. However, more work is required to explain predictions from AI systems to CHWs and other novice technology users like them.

There is a large body of research within HCI that discusses the use of text-free interfaces for novice and low-literate technology users [**? ? ?**] and ways to effectively use color in user interfaces [**? ? ? ?**]. Despite this, there is a lack of guidelines for stipulating the best use of color and visualizations in graphical approaches to XAI, leaving plenty of room for researchers to incorporate HCI principles into this domain. There is also a burgeoning amount of literature focused on examining the impact of visualizations in XAI [**? ? ?**]. However, this research shows mixed results for the ability of visual-based XAI approaches to improve user trust and understanding of model predictions. With this in mind, we call for more research into radical design approaches to visualizations within XAI.

One of the challenges with visualization-based explainability methods is the higher complexity of the information shown. While removing color from our explanations drastically improved our participants' understanding of the predictions, we cannot conjecture that colors shouldn't be used in explanations tailored to novice technology users. Additionally, given the focus of this work on just two explanation methods, it is also hard to determine if the failure is within the way *we* chose to visualize the explanations, if fundamentally the information associated with jaundice diagnosis being explained is too complicated, or if LIME and SHAP themselves are too complicated. We call for more research into how other forms of XAI that incorporate text and audio methods can be implemented to create end-user-friendly visual XAI. By leveraging these advances, along with incorporating the design recommendations presented below, we can unlock the potential for a new era of novel XAI methods that meet the needs of diverse users globally.

6 CONCLUSION

This paper details a qualitative study examining how CHWs, who are increasingly expected to work cooperatively with AI-driven tools, engage with and perceive AI explainability methods. Additionally, our work explores how XAI methods can be designed to improve CHWs' understanding of AI. Using an AI-based jaundice diagnostic app as motivation, we uncover conflicts between CHWs' X/AI understanding, mental models of disease diagnosis, color associations, and perceived benefits of having access to explanations. We conclude by discussing (1) how we can build XAI that meets user needs, (2) "scaffolding structures" to help users like CHWs understand explanations, (3) how visualizations can be leveraged to improve current XAI methods, and (4) design recommendations for XAI methods targeted toward novice AI users. Our findings highlight opportunities for new domains of XAI research to account for socially, culturally, and linguistically diverse users with low levels of AI knowledge and augment their collaboration with AI-driven tools. The existence of community health work in regions beyond India and the use of XAI in domains beyond healthcare indicates a possibility for our findings to generalize beyond CHWs in rural India. With this in mind, subsequent studies will be required to explore how this work can be relevant in other domains.

ACKNOWLEDGMENTS

We express our sincere gratitude to the participants for sharing their insights. We are also grateful to NYST for their facilitation of our research. This work was funded in part by the National Science Foundation grant #1748903 and the Cornell Center for Health Equity.

Received January 2023; revised October 2023; accepted December 2023