Making AI Explainable in the Global South: A Systematic Review

Chinasa T. Okolo
Computer Science
Cornell University
Ithaca, New York, United States
chinasa@cs.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

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) are quickly becoming pervasive in ways that impact the lives of all humans across the globe. In an effort to make otherwise "black box" AI/ML systems more understandable, the field of Explainable AI (XAI) has arisen with the goal of developing algorithms, toolkits, frameworks, and other techniques that enable people to comprehend, trust, and manage AI systems. However, although XAI is a rapidly growing area of research, most of the work has focused on contexts in the Global North, and little is known about if or how XAI techniques have been designed, deployed, or tested with communities in the Global South. This gap is concerning, especially in light of rapidly growing enthusiasm from governments, companies, and academics to use AI/ML to "solve" problems in the Global South. Our paper contributes the first systematic review of XAI research in the Global South, providing an early look at emerging work in the space. We identified 16 papers from 15 different venues that targeted a wide range of application domains. All of the papers were published in the last three years. Of the 16 papers, 13 focused on applying a technical XAI method, all of which involved the use of (at least some) data that was local to the context. However, only three papers engaged with or involved humans in the work, and only one attempted to deploy their XAI system with target users. We close by reflecting on the current state of XAI research in the Global South, discussing data and model considerations for building and deploying XAI systems in these regions, and highlighting the need for human-centered approaches to XAI in the Global South.

CCS CONCEPTS

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

KEYWORDS

Artificial Intelligence, Machine Learning, Algorithmic Fairness, Explainability, HCI4D, XAI4D, ICTD, Global South

ACM Reference Format:
Chinasa T. Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI Explainable in the Global South: A Systematic Review. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '22), June 29-July 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3530190.3534802

COMPASS '22, June 29-July 1, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9347-8/22/06…$15.00
https://doi.org/10.1145/3530190.3534802

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) are quickly becoming pervasive in many aspects of our lives. These techniques are now impacting the ways humans are hired [76, 150], receive credit scores [78], shop for products [4, 116], and receive social assistance [5, 77, 106]. Despite the ability of AI to work for the greater good, research has shown the negative impacts of employing such technologies. For example, studies have revealed gender bias in facial recognition systems [22], racial bias in recidivism prediction [65, 94], and ethical issues in automated hiring [76, 150], credit scoring [78], and surveillance [60, 152].

In recognition of the growing influence of AI/ML in people's lives and the problems around many AI/ML systems being "black boxes", a growing body of work is focusing on "Explainable AI" [29, 80, 86, 128, 129]. We consider "Explainable AI" to be the set of machine learning techniques (algorithms, toolkits, frameworks, etc.) that equip humans involved in the design, development, integration, and use of AI-enabled systems with the ability to comprehend, trust, and manage them [10]. However, existing work in XAI is primarily focused on contexts/communities in the West/Global North. But, AI/ML is increasingly also pervading contexts in the Global South with goals to improve healthcare [16, 96, 124], aid in social policy decision-making [54, 81, 102], improve educational access [21, 91], and address environmental concerns [68, 74, 141, 143], with potentially life-changing consequences for individuals and communities [15]. Thus, there is an urgent need to also closely examine the current state of XAI in the Global South, how AI/ML systems are being made to be understandable by people who must use these systems, and identify opportunities to make AI/ML systems more explainable for users in these contexts.

While several scholars have analyzed emerging trends in research related to XAI [2, 10, 47, 52, 101], these respective works do not specifically focus on the Global South, our target region. In this paper, we contribute a systematic review of XAI research in the Global South both for researchers working in the field of XAI and the larger ICTD community. Our search for literature yielded 16 papers focused on XAI in the Global South. We then analyzed these 16 papers to understand factors that shape the design and deployment of XAI systems for these regions. Although we did not limit the date range in our search, all 16 papers in our review were published within the last three years (2019-2021), highlighting the nascent and emerging nature of this field. The 16 papers were published at 15 different venues including COMPASS, CHI, and workshops at premier ML conferences like the Conference on Neural Information Processing Systems (NeurIPS), the International

Conference on Learning Representations (ICLR) and the ACM Conference on Knowledge Discovery and Data Mining (KDD). We show how, to date, XAI research in the Global South centers mainly on India, with little work focusing on Africa or Latin America. In the papers we reviewed, there was a strong focus on applying XAI to socially-relevant domains, with healthcare being the most popular domain. The papers also studied a diverse range of communities, from sex workers to secondary school educators to fashion retailers. Of the 16 papers, 13 focused on technical implementations of XAI, all of which used at least some data that was local to the context they worked in. However, only three of the 16 papers engaged with intended target communities, and only one actually deployed their XAI system with real users.

Taken together, our findings reveal both encouraging trends and concerning gaps emerging in XAI research in the Global South. We close by discussing important data and model considerations that AI and ICTD researchers need to take into account as they work to design and build fair, ethical XAI tools for contexts and communities in the Global South. Finally, we call for more human-centered approaches to XAI research to ensure that the field addresses growing inequities and makes AI explainable to people from diverse, global contexts.

2 BACKGROUND AND RELATED WORK

2.1 AI/ML in the Global South

Although our review focuses specifically on XAI, we situate our work within the broader context of prior and ongoing AI/ML research in the Global South. A growing body of work has began to emerge exploring the potential of AI and its respective benefits and implications in the Global South. For example, work from Dufresne-Camaro et al. [34] focuses specifically on computer vision, surveying the field to identify risks and recommendations for practitioners building computer vision systems for this region. Other work has examined AI initiatives on national, regional, and continental levels within the African context, providing an overview of the the benefits and challenges resulting from AI use in this region [53].

Although the concept of "responsible AI" has primarily been discussed in relation to contexts in the Global North [10, 28, 118], research has begun to explore how these concepts translate to developing regions. For example, to understand how concepts of algorithmic fairness differ from Western contexts, Sambasivan et al. [126] analyzed challenges surrounding ML deployments in India. From this analysis, the authors present a roadmap for improving fairness in India with actionable steps, such as considering infrastructure limitations, problem choices, dataset building, and model training.

Scholars have also focused on providing recommendations geared towards governments to help inform how AI can be leveraged towards sustainable development goals (SDGs) [123, 145, 147]. For example, Isagah and Musabila [61] interviewed ML practitioners across the African continent to understand their experiences developing AI/ML systems, implementation challenges associated with system deployments, and provide policy recommendations to African governments. As AI continues to be actualized within the Global South, it will take collaborative efforts to ensure the fair development and use of these systems.

Our paper expands prior research on AI/ML in the Global South by focusing specifically on XAI, a new subarea of the field that has to date been underexplored in low-resource contexts. Our analysis aims to highlight ways in which XAI might contribute to the development of ethical, understandable, and equitable AI/ML systems in the Global South.

2.2 Explainable AI

Recent years have seen a rapid increase in research addressing the sociotechnical issues brought about by the use of AI systems in the Global North. In particular, the field of XAI has arisen, with the goal of ensuring that the decisions and recommendations made by AI systems are understandable to people who interact with such systems [10]. Explainability can focus on both the model design and output in regards to how humans understand decisions made by AI systems. Incorporating explainability into model design and development allows developers to understand how model parameters are making decisions and how these decisions are tracked [117]. When explaining ML models, the foremost way to do so is through techniques such as feature importance [119, 120] and model distillation [79, 85, 140], methods that are not accessible to those who lack such specialized knowledge. In the development of XAI systems, techniques such as looking at the prediction accuracy, limiting the scope of decision-making, and educating teams working with AI on these systems can help improve how developers understand the decisions made by their respective models [58]. Explainability has become a prominent issue, especially for corporate adopters of AI who need to ensure that decisions made by their AI systems can both be explained and understood [113]. We now discuss prevalent XAI techniques, vocabulary, and concepts, with the goal of providing a strong basis for understanding the terms used in our search and filtering methodologies, and the techniques used within the our resulting corpus.

XAI Techniques. Popular techniques for XAI include SHAP, a method that explains individual predictions by measuring how model features contribute to them [88], and LIME which trains surrogate models to help explain how the original model arrived at a particular decision [119]. SHAP and LIME methods are relatively generalizable and have been applied to a wide range of machine learning subfields like computer vision and natural language processing (NLP). In computer vision, these techniques have been leveraged to develop visual explanation maps [67, 82] and saliency approaches [97] that highlight regions of interest in images. Anchor, a method that explains individual predictions of classification models, works for text or tabular data [120]. Other methods from philosophical domains include contrastive explanations [33, 64] which aim to describe event occurrences in contrast to another and counterfactuals [24, 51, 70] which describe events in causal form.

Over the years, companies such as Microsoft [92], IBM [12], Google [49, 50], and Amazon [8] have developed commercial toolkits for use in developing ML systems with their respective cloud platforms. Many of these toolkits incorporate popular XAI algorithms such as LIME and SHAP or create improved versions of these algorithms [43] while providing features such as tutorials

and visualizations to ease their integration by developers. Software libraries have also become popular tools for developers to rely on to improve the explainability of models. Tools such as Alibi [75], Skater [32], InterpretML [105], EthicalML-XAI [42], DALEX [20], and iNNvestigate [6] provide explainability in a variety of ways, ranging from model inspection to prediction explanation. Many of these techniques and tools focus on software developers, rather than end users, and require a relatively high level of AI/ML knowledge to implement.

Human-Centered Studies in XAI. While much of the research presented in XAI has focused on technical and algorithmic techniques, human-centered studies have begun to emerge that aim to understand how humans perceive, value, and understand AI systems. Ehsan et al. [35] introduced the concept of "human-centered explainable AI (HCXAI)" with the aim of centering humans in the design of AI systems while understanding the unique factors that shape each respective user. Work from Katell et al. [69] focused on the use of participatory and co-design methods to ensure algorithmic accountability in ML interventions deployed in local communities. Liao et al. [83] interviewed user experience (UX) and design practitioners with the aim of identifying gaps in existing XAI research and opportunities to create XAI tools that make AI understandable for users in real-world settings.

To help non-technical users understand AI systems, researchers have explored techniques such as interactive demonstrations [48, 136], visualizations [99, 138, 139], and storytelling [37]. For example, Hsu et al. [55] build on prior work centering humans in XAI research by leveraging participatory design practices to encourage co-design of AI systems between scientists and local communities. Cheng et al. [27] conducted online experiments with non-AI experts and find that explanations improve user understanding of algorithms but come with time tradeoff. Additionally, frameworks that guide the design, development, and use of XAI techniques have also become more common [40, 149, 155]. For example, work by Alikhademi et al. [7] aims to evaluate explainable AI tools in terms of how well they explain results to the users of AI systems.

Taken together, the growing literature on human-centered XAI suggests an exciting focus on work that makes AI understandable to non-technical users. However, all of this work has been done in the Global North. As a result, very little is known about the state of XAI that focuses on contexts in the Global South. This gap is concerning, given the growing enthusiasm for deploying AI/ML systems in communities in the Global South [11, 30]. Our paper thus provides a first look at early research emerging in this important area.

3 METHODS

We conducted a systematic literature review to identify XAI research focused on contexts in the Global South. We now outline the methodology for our review, including the venues, search terms, and filtering strategies used. We then discuss how we analyzed and categorized the 16 resulting papers.

Search Methodology. To identify literature for inclusion, we incorporated methods similar to researchers conducting other systematic reviews [31, 63, 144]. Over a period of three months from August 2021 to October 2021, we surveyed literature to identify papers that discuss XAI techniques applied in the development, deployment, and evaluation of AI-enabled technologies in the Global South. We searched the ACM Digital Library, arXiv, and Google Scholar, utilizing their built-in search functions. Our search was not restricted to any specific time frame and not limited to specific venues or proceedings. Our searches yielded papers across a range of venues within AI and HCI including conferences (e.g., CHI and COMPASS) and workshops (e.g., Machine Learning for Development Workshop at NeurIPS, AI for Social Good Workshop at ICLR, and Responsible AI Workshop at KDD).

Our search combined two sets of search terms (see Table 1): (1) terms in the category of *XAI*, including phrases such as 'AI explainability', 'model explainability', and 'explainable AI', and (2) terms in the category of *the Global South*, including phrases such as 'developing countries', 'developing regions', and 'developing world'. To create a combined term, we took one phrase from the XAI category and another from the Global South, and used the boolean operator 'AND' to ensure that resulting papers included both phrases. The combination of search terms from both categories resulted in 192 search terms.

Searches from the ACM Digital Library were configured to filter out demonstrations, extended abstracts, posters, talks, and tutorials. However, arXiv and Google Scholar did not have similar filtering functions, so we manually filtered the search results to maintain this criteria. Our initial searches from the terms listed in Table 1 yielded 208 papers.

We then conducted multiple passes over the search results. Due to the similarity of many of our search terms, we prioritized removing duplicates in our first pass. We also removed papers where the search terms appeared only in the references section of the paper or the paper was a review or survey of XAI techniques (survey papers were removed since they talked generally about XAI and did not focus specifically on work that was conducted in the Global South).

After these passes, we were left with 115 papers in our shortlist. Next, we manually filtered these 115 papers to identify those that: (1) discussed the benefits or implications of XAI, or (2) described the design, development, evaluation or implementation of an AI system or machine learning model that incorporates XAI methods. We reviewed each paper by carefully reading the title, abstract, and introduction, and skimming the other sections. We used this methodology to remove papers that, upon closer inspection, did not focus on either XAI or its applications to the Global South. An example of a paper that we filtered out was *"Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment* by Wang et al. [148]. While the keywords initially seemed relevant to our search, upon closer inspection, we found that the paper focused on challenges regarding the deployment of a clinical support tool while only mentioning in one sentence the possibilities of integrating XAI in future iterations of the tool. At the end of this phase, we were left with a final set of 16 papers from 15 venues (see Table 2).

Analysis. After finalizing our dataset of 16 papers, we analyzed the content of each paper. We began familiarizing ourselves with the papers by carefully reading through them. We developed a rubric based on key information extracted from each paper (discussed

Table 1: Strategy used to compose search terms

Category	Keywords
(A) Explainable AI	AI explainability, model explainability, explainable AI, explainable artificial intelligence, explainable ML, explainable machine learning, algorithmic explicability, algorithmic explainability
(B) Global South	Africa, community health workers, developing countries, developing regions, developing world, economic development, farmers, global development, global south, India, international development, Kenya, Latin America, LMIC, low-income, low-literate, low-resource, marginalized, novice technology users, rural, social good, South Africa, Southeast Asia, underserved
Search Term	(A) AND (B)

below) and entered this information into a document management system. Multiple passes were done by the team to ensure that our categorizations were accurate. Additionally, we held regular meetings to discuss discrepancies and reach consensus on the final categorizations.

For each paper, we extracted its geographic focus (e.g., Africa, Latin America, India, China), domain (e.g., healthcare, education, government), population involved (e.g., indigenous populations, community health workers, refugees), the methods used (e.g. algorithm development, dataset building, interviews), and whether the technology, prototype, or interviews described in the paper were implemented or conducted with local populations. We use the term "implemented" to describe an XAI tool or system that is brought into use in local communities. An example of a paper that does not implement an XAI system is "Learning Explainable Interventions to Mitigate HIV Transmission in Sex Workers Across Five States in India" by Awasthi et al. [13]. In this paper, the authors describe their work building a Bayesian network combined with an XAI model to encourage safe sex practices with sex workers in India. At the time of publication, the paper was not tested in a real-world setting but the authors state that the model is *"currently ongoing field trials for assessing the real-world utility of this approach."*

For papers that surveyed participants or included data from more than one country, we either categorized this as "Global South" broadly or, if localized to a continent, we used the respective name of the continent (e.g., "Africa"). To define domain, we described it as the primary field in which the paper describes or addresses a problem. An initial list of domains was obtained from prior meta-analysis papers focusing on human-computer interaction for development (HCI4D) [31]. However, we expanded this list upon closer analysis of the papers in our dataset, adding domain areas such as "Government/Policy", "Finance", "ML/Algorithmic Development", "Informational/Awareness", and "Misinformation" to existing topic areas, such as "Healthcare", "Education", and "Agriculture". Next, we categorized the papers around the types of XAI techniques used, engagement with end users, and whether the XAI systems were deployed. Finally, we analyzed the papers to distill trends in XAI research in the Global South and the implications, benefits, or challenges encountered in the research.

Positionality. All authors are from countries in the Global South, currently residing in the US. All have extensive experience conducting fieldwork with underserved communities in different continents. We believe in the importance of centering humans in AI and XAI research. We view research from an emancipatory action research mindset [14, 73], aiming to highlight the opportunities, challenges, and tensions of incorporating XAI techniques into the design of systems for communities in the Global South.

Limitations. The goal of our paper is to broadly survey the state of XAI in the Global South. We acknowledge that our searches were limited to papers written in English and indexed by the scholarly platforms used in our literature review. We realize that scholarly work is not limited to prestigious journals and conferences and made an effort to also searching archival preprint platforms like arXiv and include workshop papers. Despite these limitations, our analysis provides useful insights to inform future work shaping XAI advances in the Global South.

4 FINDINGS

The 16 papers in our dataset (listed in Table 2) cover a broad range of domain areas, publication venues, and regions, and engage with a diverse range of populations. We start by discussing when and where the papers were published, followed by the countries and geographic regions involved in the work. We then turn our attention to the domains the papers focus on, the populations studied, and whether the data used in each paper was local to the context. Next, we discuss whether or not the research involved target users and the specific XAI methods employed. Finally, we look at the challenges that arose when putting XAI tools into practice.

When and where were the papers published? Although our search did not specify a date range, all 16 papers in our dataset were published within the last three years: one in 2019, nine in 2020, and six in 2021. This finding suggests that research on XAI in the Global South is very new. We further note that the total number of papers in our review is very small (16) compared to the overall amount of XAI research published over the same timeframe of three years (approx. 18,000 papers according to Google Scholar). These findings highlight the timeliness of our review, which provides a first look at research emerging in the space, laying a foundation for understanding the state of XAI in the Global South.

The 16 papers were published in 15 different venues, including journals (6), workshops (5), and conferences (4), with one paper published as a preprint on arXiv. The large number of different venues suggests that research in this space has not yet found an academic 'home' venue (e.g., only one paper was published at COMPASS and another one at CHI). Given the diverse range of potential users and communities that have been targeted by the papers within our corpus, the formation of a venue that can cater to the needs of researchers focusing on XAI in the Global South may be necessary. In addition, the relatively large proportion of workshop papers may be helped by several recent workshops specifically focused on

Table 2: Final set of 16 papers obtained from our filtering strategy

Title	Authors	Venue	Year
Working women and caste in India: A study of social disadvantage using feature attribution	Kuhu Joshi and Chaitanya K. Joshi	AI for Social Good Workshop at the International Conference on Learning Representations (ICLR)	2019
Automating the Surveillance of Mosquito Vectors from Trapped Specimens Using Computer Vision Techniques	Mona Minakshi, Pratool Bharti, Willie B. McClinton, Jamshidbek Mirzakhalov, Ryan M. Carney, and Sriram Chellappan.	ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)	2020
Detecting and Explaining Depression in Social Media Text with Machine Learning	Rida Zainab and Rajarathnam Chandramouli	Designing AI in Support of Good Mental Health Workshop, ACM Conference on Knowledge Discovery & Data Mining (KDD)	2020
Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva	Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen.	IEEE Symposium Series on Computational Intelligence (SSCI)	2020
Improving healthcare access management by predicting patient no-show behaviour	David Barrera Ferro, Sally Brailsford, Cristián Bravo, and Honora Smith	Decision Support Systems	2020
Interpretable Poverty Mapping using Social Media Data, Satellite Images, and Geospatial Information	Chiara Ledesma, Oshean Lee Garonita, Lorenzo Jaime Flores, Isabelle Tingzon, and Danielle Dalisay	Machine Learning for Development Workshop (ML4D) at the Conference on Neural Information Processing Systems NeurIPS	2020
Learning Explainable Interventions to Mitigate HIV Transmission in Sex Workers Across Five States in India	Raghav Awasthi, Prachi Patel, Vineet Joshi, Shama Karkal, and Tavpritesh Sethi	Machine Learning for Development Workshop (ML4D) at the Conference on Neural Information Processing Systems NeurIPS	2020
Socio-ethical implications of using AI in accelerating SDG3 in Least Developed Countries	Kutoma Wakunuma, Tilimbe Jiya, and Suleiman Aliyu	Journal of Responsible Technology	2020
Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors	Carlos Eduardo Beluzo, Everton Silva, Luciana Correia Alves, Rodrigo Campos Bresan, Natália Martins Arruda, Ricardo Sovat, and Tiago Carvalho	Informatics in Medicine Unlocked	2020
The use of machine learning "black boxes" explanation systems to improve the quality of school education	Ravil Muhamedyev, Kirill Yakunin, Yan A. Kuchin, Adilkhan Symagulov, Timur Buldybayev, Sanzhar Murzakhmetov, and Alibek Abdurazakov	Cogent Engineering	2020
Applying a principle of explicability to AI research in Africa: should we do it?	Mary Carman and Benjamin Rosman	Ethics and Information Technology	2021
An Empirical Study of Accuracy, Fairness, Explainability, Distributional Robustness, and Adversarial Robustness	Moninder Singh, Gevorg Ghalachyan, Kush R. Varshney, and Reginald E. Bryant	Responsible AI Workshop at ACM Conference on Knowledge Discovery and Data Mining (KDD)	2021
Explainable AI based Interventions for Pre-season Decision Making in Fashion Retail	Shravan Sajja, Nupur Aggarwal, Sumanta Mukherjee, Kushagra Manglik, Satyam Dwivedi, and Vikas Raykar	ACM IKDD CODS, COMAD (CODS COMAD)	2021
Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI	Indranil Ghosh and Manas K. Sanyal	International Journal of Information Management Data Insights	2021
"It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India	Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha	ACM Conference on Human Factors in Computing Systems (CHI)	2021
Tracking Peaceful Tractors on Social Media–XAI-enabled analysis of Red Fort Riots	Ajay Agarwal	arXiv	2021

related topics, such as the Machine Learning for the Development workshop at NeurIPS. We hope that such workshops might eventually serve as a gateway for a larger number of full papers accepted at the corresponding conferences, although our dataset does not yet show evidence of this.

What countries and geographic regions are represented? Next, we analyzed the locations where the research took place. Prior reviews of ICTD research has shown a skew towards work being conducted in India [31, 111, 144]. Despite prior work calling for more geographic diversity [31], we found similar trends in our analysis (see Table 3). Six papers focused on XAI in India, followed by five focused on the broader Global South (meaning that the research covered multiple countries in this region). From there, papers focused on African continent (1) and countries like Brazil (1), Colombia (1), the Philippines (1), and Kazakhstan (1). The relatively large proportion of papers focused on India is likely due to the presence of a large number of academic institutions as well as industry research labs like that of Microsoft, IBM, and Google. We also find it important to note the small amount of work that focused on Africa, the Caribbean, and Central and South America. In regions where English is not the primary language for academic research, it is possible that scholarly work appeared in local languages (e.g., Spanish) and this could be the reason for the lack of work we found focusing on Central and South America. However, the small number of papers focusing on Africa is perhaps surprising, given the existence of organizations such as Deep Learning Indaba (established in 2017) [59], Data Science Nigeria (established in 2016)

Country/Region	No.
India	6
Global South	5
Africa	1
Brazil	1
Colombia	1
Philippines	1
Kazakhstan	1
Total	16

Focus Area	No.
Healthcare	6
Government & Policy	3
Finance	2
General Awareness	1
Education	1
Misinformation	1
Retail	1
Algorithm Development	1
Total	16

Publication Type	No.
Journal	6
Workshop	5
Conference	4
Pre-print	1
Total	16

Table 3: Summary of the 16 papers in our review by geographic region/country, focus area, and publication type.

[103], and Masakhane NLP (established in 2019) [109], all of which aim to increase AI research and practice in Africa.

After understanding what countries and geographic regions were represented in our dataset, we were interested in examining where the authors were based versus where their respective research took place. Our analysis showed that authors came from institutions based in India, Brazil, South Africa, Kenya, the Philippines, Kazakhstan, Colombia, Armenia, Belgium, the United States, and the United Kingdom. Encouragingly, most papers were produced by institutions in the Global South. When mapping the country the work focused on versus the authors' institutions, it was rare if at least some of the authors were not based in the country. For example, five out of six papers that focused on India were published by authors based in institutions in India. In many cases, the research team comprised of members located both in the Global South and the Global North. The paper by Okolo et al. [108] is an example of such work where research team consisted of members based in India as well as the United States. Other examples of such heterogeneity include work by Ferro et al. [41], where researchers had multiple affiliations with institutions in the Global North and South (e.g. Colombia and the United Kingdom).

What are the domain areas that papers focus on? Next, we analyzed the domain areas that motivate researchers to explore the feasibility of XAI methods and techniques. We found that papers targeted a range of domains (see Table 3), including algorithmic development, healthcare, government & policy, education, misinformation, retail, and finance.

A substantial number of papers (6) focused on healthcare. For example, Beluzo et al. [17] focused on predicting neonatal mortality, while Ferro et al. [41] worked on estimating patient no-show behaviors. Awasthi et al. [13] used survey data collected from over 10,000 female sex workers in India to develop an ML model to understand the factors that influence their safe sex practices. In doing so, they described how developing *"a transparent and explainable approach with expert evaluation is critical to model sensitive issues"* and implied that their work could be put into practice for similar interventions more generally. Several of these papers cite the lack of quality healthcare services in low-resource areas as a primary motivator for applying AI and XAI to problems in healthcare [13, 17, 93]. In line with this general motivation, prior work in the Global North that has focused on AI more broadly (as opposed to our focus on XAI in particular), has also pushed to develop AI-enabled tools for use within healthcare, such as improving the delivery of primary care services while preserving the autonomy of healthcare practitioners and maintaining the value of human presence [44, 71, 115].

Addressing socioeconomic concerns to aid in policy and government decision-making was also a strong focus for three papers in our dataset. For example, work by Joshi and Joshi [66] examined the caste-group structure in India to understand changes in the ability to predict the work status of women across those castes. By applying an XAI technique (SHAP) to their model, the authors found that over time, caste affiliation is less likely to determine working status (not/having a job) in younger generations of women and that these women are more likely to experience upward mobility in their careers. This work shows promise in improving how social stigma is measured in a longitudinal manner and could shape programs geared towards improving economic access for lower status castes in India.

In another paper that focused on making policy recommendations for reducing poverty, Ledesema et al. [81] use government survey data, social media data, and geospatial features to map poverty in the Philippines. Their approach overcomes challenges associated with traditional poverty mapping by using lower-cost data instead of high-resolution satellite images. By leveraging XAI algorithms to help quantify the impact of dataset features on predictions, the team presents a feasible approach that could find applications in other research domains within the Global South. In situations where policymakers and government officials may not possess the capacity necessary to understand the decisions made by algorithmic systems, incorporating XAI methods in the development of such systems could improve how AI is used to address social issues.

Who are the studied populations? Research in the fields of HCI4D and ICTD strongly emphasize paying attention to local contexts when designing interventions for communities. Drawing on these lessons, we aimed to analyze what communities or populations researchers identified as the intended beneficiaries of their work and/or who contributed to their research (e.g., by providing data).

To develop or analyze XAI tools for applications in the healthcare domain, scholars focused on women sex workers [13], community healthcare workers in rural India [108], local communities seeking healthcare services in Colombia [41], taxonomists [93], social media users who speak Urdu [154], and mothers and newborns in Brazil [17]. In education, one paper focused on students and educators in secondary school settings [98]. For misinformation, one paper focused on tweets on sociopolitical events in India [3]. Within papers that focused on government and policy issues, populations included were women in India from a range of castes [66], households in the Philippines [81], borrowers on a loan request platform [137], households from four East African countries [135],

and product developers and designers in the retail fashion industry in India [125].

Several papers did not mention a specific population in their analysis, focusing instead on analyzing XAI for a general population in the Global South [26, 46, 147]. An example of this is work by Ghosh and Sanyal [46] that uses financial market data from India to create a model predicting market volatility during the COVID-19 pandemic. Overall, we found the diverse populations targeted by the papers in our dataset encouraging, suggesting a diversity of potential applications of XAI across many different settings in the Global South.

Is the data used local to the context? A growing body of AI/ML research has called for ethical collection and use of data used to train models for intelligent systems [22, 57, 112, 127, 131]. Situating this within the context of HCI4D and ICTD research, work in XAI in the Global South should prioritize using data local to and representative of the populations that will be affected by such systems. When framing the relevance of the problems addressed within the papers we analyzed, we were interested in whether the data used was local to the context addressed. Encouragingly, we found that, of the 13 papers that focused on technical implementations of XAI, all used data from regions in the Global South. For example, Singh et al. [135] conducted an empirical study to analyze different metrics across the fairness, accountability, transparency, and ethics (FATE) spectrum, including explainability, on a series of eight datasets. Two out of the eight datasets consisted of data from regions in the Global South: the first is a dataset containing demographic and poverty-level features for over 70,000 households in Mexico and the second contains data describing demographics of and the use of financial services by over 33,000 people in Kenya, Rwanda, Tanzania, and Uganda.

Prior work has highlighted the lack of relevant training data available for contexts in the Global South [1, 132]. For the papers in our dataset, when relevant data was not available, researchers commonly built their own datasets to fill this gap. For instance, Beluzo et al. [17] constructed a dataset with over one million features, extracting data from birth and death systems in Brazil to understand the socioeconomic and demographic factors contributing to neonatal deaths. Agarwal et al. [3] created a dataset composed of 50,000 tweets posted before, during, and after a sociopolitical event that occurred in India. They then leveraged this dataset to create an XAI tool to characterize the spread of disinformation and misinformation on social media. Similarly, Zainab and Chandramouli [154] created a dataset of over 16,000 Reddit posts and developed an XAI model to understand what features in social media posts indicate depression. Such work highlights the additional labor required in generating local datasets that researchers in the Global South may need to undertake when implementing XAI systems.

Did the research involve target users? Human-centered design and HCI4D best practices generally require that researchers engage with intended users, often via fieldwork to understand the needs of these users before building or deploying new technologies [31]. Therefore, we wanted to see if papers in our dataset followed these basic best practices. Only three of the 16 papers engaged with target users [13, 108, 125]. In one of these papers, Sajja et al. [125], researchers at IBM Research India, partnered with a local fashion house to develop an XAI tool for forecasting sales of clothing products. They first created personas to understand the two types of users of their tool: clothing designers and commercial buyers (individuals who determine what items are stocked in retail stores). Next, the researchers actively consulted with the users to inform their design decisions for the XAI tool and used techniques such as SHAP and counterfactuals to explain predictions for product recommendations. Finally, in a survey used to gauge the usefulness of the developed tool, the researchers obtained feedback from their users, finding that the tool improved their respective workflows and made it easier to understand sales trends. We find it particularly notable that the product was built with the end users in mind and that the developers were intentional in including them throughout the entire development process.

Instead of directly engaging end users, in one case, researchers worked with domain experts to design an XAI tool to mitigate the impact of mosquito-borne diseases like malaria, dengue, West Nile virus and Zika fever. Minakshi et al. [93] consulted with taxonomists in Uganda, India, Brazil, and the United States to develop a tool to help public health workers identify mosquitoes capable of spreading diseases. The authors incorporated explainability into their tool by visualizing feature maps that highlighted pixels within the output images that were most significant in classifying the genus or species of mosquitoes. While taxonomists were not necessarily the target users for this tool, the authors used feedback from these experts to validate the results from the ML model, which they plan to deploy soon with public health partners in low-resource regions. In addition to the technical expertise, the researchers received affirmation from taxonomists who emphasized the need to design such a tool. In this work, the researchers shared: *"The taxonomists we partnered with in India, Sub-Saharan Africa, Brazil and USA are all close to 70 yrs old, and they indicated that taxonomy is indeed a dying field, and related expertise is hard to attract and train."* In cases like the problem presented by Minakshi et al. where an XAI tool is being introduced with the intent of replacing skilled experts, the knowledge of such experts is still needed to ensure that the results produced from XAI systems are both accurate and explainable to end users who may not possess specific domain knowledge.

Unfortunately, our analysis shows that these positive examples of engaging with intended users or other stakeholders was relatively rare (only 3 out of 16 papers). This is concerning given the generally agreed upon importance of tailoring solutions to local contexts and communities in HCI4D and ICTD research.

What XAI methods were used? Since the inception of XAI, methods have shifted from being purely technical to incorporating human-centered techniques that aim to make AI understandable to non-experts and users in real-world settings [38, 69, 84, 99**?**]. Thus, we wanted to understand what methods and techniques were being employed by XAI work targeting the Global South. The majority (13) of the papers in our dataset used a technical XAI method to develop an ML model or to conduct an analysis on data. We considered a paper to be "technical" if it developed or used an XAI algorithm, toolkit, or library such as the ones described in Section 2.2. Nearly all of the papers used an algorithm such as SHAP or LIME to understand how features within their datasets contribute to

model decision-making. An example is work by Stevens et al. [137] who used SHAP to explain the process of predicting loan outcomes on Kiva.org, a website that crowdfunds loans for entrepreneurs in developing countries. Their work also incorporated the use of AI Fairness 360, a toolkit developed by IBM [58], which uncovered a potential gender bias on the platform.

While the existing XAI algorithms and libraries are generally helpful in illuminating decisions made by ML models, these explanations often require a significant amount of technical knowledge to both implement and interpret them. Visual-based XAI methods could be a step towards centering explainability around novice technology users and others with low AI/ML domain knowledge. An example along these lines is work by Minakshi et al. [93] which used XAI to produce visualizations highlighting pixels within images processed by their AI model that indicate features of specific genus or species of mosquitoes.

Several researchers incorporated multiple XAI methods into their model development [46, 125, 137]. For example, in work using XAI to understand what factors influenced market volatility during the first year of the COVID-19 pandemic in India, Ghosh and Sanyal [46] use the Shapash library to conduct their analysis which leverages two XAI techniques: LIME and SHAP. While more explanations may be useful to help improve model comprehension, the usage of XAI also brings up concerns around the increased computational complexity of AI models [56] and high computing resources needed to run them.

Of the three papers that did not use technical XAI methods, we found that two used qualitative research to explore XAI. For example, Okolo et al. [108] conducted an interview-based qualitative study in rural India to examine community health workers' perceptions of how AI-enabled tools would fit into their workflows. The researchers used insights from the interviews to examine how designers and developers of AI systems can leverage explainability for novice technology users, especially those situated in low-resource environments within the Global South. Similarly, Wakunuma et al. [147] examined the potential of AI in advancing the third Sustainable Development Goal (SDG) that focuses on good health and well-being. The researchers employed a SWOT (strengths, weaknesses, opportunities, and threats) analysis—a method commonly used in strategic planning—to explore the socioethical implications involved with employing AI to address the third sustainable development goal and how explainable AI can be used to improve model transparency in medical decision-making.

The last non-technical paper in our dataset focuses on critically examining the use of XAI to address global development challenges in Africa. In their work, Carman and Rosman [26] use epistemological framing (by asking the question 'how does it work?') to survey the development of XAI and examine the feasibility of applying XAI within the African context. Such framing aims to understand how XAI can work to be in alignment with African values and interests. When applying explainability to AI research using an epistemological framework, the authors propose identifying objectives and goals for intelligent systems that align with the local contexts they will operate in. This goal aligns directly with common practices in HCI4D and ICTD that aim to ensure technologies are developed *for* and *with* local communities. Carman and Rosman also acknowledged the onerous demand that African researchers may face when integrating explainability techniques, tentatively proposing a division of labor to address such concerns. To anticipate ethical issues that may occur when developing XAI systems for use within Africa, Carman and Rosman propose that ML practitioners work in tandem with experts in the social sciences and humanities to ensure that the underlying values of respective societies are prioritized and embedded in the resulting systems.

What challenges arose when building and putting XAI tools into practice? Finally, well-known best practices in ICTD often call for deploying technologies directly with communities [9, 23]. However, within the field of AI in general and XAI in particular, we find this to be a rare occurrence. Only one of the 16 papers in our dataset developed an XAI system that was put into practice with real-world users [125]. This small number indicates that deploying XAI systems is still a novel practice and many of the papers in our dataset were exploratory in nature, and not at the deployment stage.

Next, we were interested to see what challenges and issues were encountered as the researchers tried to put XAI into practice. In their work, Sajja et al. [125] developed an XAI tool to forecast demand for products in the fashion retail industry in India. The researchers collaborated with a local fashion house, deploying the system to analyze past fashion seasons and to help plan for future ones. The researchers noted data being a challenge at multiple points in the development lifecycle of their XAI model. First, the researchers mention the limited number of data points that prevented them from building an explainable model at a level comparable to the baseline modeling framework, indicating the extra work needed to gather additional data and to employ XAI techniques in a real-world setting. Prior research has noted the lack of "data equity" in the Global South, detailing the extreme efforts needed by AI/ML practitioners in this region to collect, curate, and employ data for use due to the lack of access to relevant datasets and limited infrastructure to train models [127]. Building upon this point, Sajja et al. also highlight the extra effort undertaken by the designers they collaborated with to contribute detailed data to the system before operating it. Several other pieces of scholarly work have detailed similar challenges when collecting data for algorithmic systems in developing contexts, noting the burden data collection and processing takes on frontline healthcare workers, data stewards, and ML developers in the Global South [62, 142]. As one of the few XAI systems deployed for real-world use in the Global South, Sajja et al.'s work provides a much needed perspective into the realities and challenges of deploying XAI systems into the wild.

The overall lack of deployments seen within our dataset is a bit worrying, but falls in line with the majority of AI systems not being deployed for real-world use. We also understand that there exist a combination of challenges in the Global South including infrastructural limitations, lack of AI practitioners, and a large number of novice technology users that may possibly inhibit the successful integration of such systems, as seen in work by Beede et al. [15]. Despite the lack of deployments, a few papers in our dataset mentioned the possibility for future work to implement their respective systems. For example, Beluzo et al. [17] state: *"we intend to evaluate the applicability of the proposed model in Brazilian data."* Another paper mentions how *"these insights have led to a*

currently ongoing field trial for assessing the real-world utility of this approach" [13]. Minakshi et al. [93] also mention *"currently, we are focusing on piloting our system in low income countries where taxonomic expertise is harder to find."* Similarly, Agarwal et al. [3] note the ongoing development of their XAI-enabled classifier to measure misinformation spread during socio-political events in India.

Progress towards putting XAI systems into practice is an important next step in testing the viability of such technologies. As XAI development continues to progress within the Global South, we hope to see real-world deployments rise in number. However, if researchers deploy XAI techniques that do not work, it might cause more harm than good to local populations. With this in mind, cautious approaches to deployment are needed when putting nebulous technologies like XAI into practice.

5 DISCUSSION

In recent years, there has been growing enthusiasm from governments, companies, and academics about the potential for AI to solve important problems in the Global South, including in healthcare [25, 100, 110, 121, 134, 151], poverty [54, 81, 102], agriculture [87, 130], and other high-stakes domains. However, research has shown that AI also has the potential to exacerbate and reinforce systemic problems, including bias and discrimination [19, 39, 104]. If care is not taken to make AI systems explainable and understandable to the people who will use them, they may end up causing more harm, particularly to marginalized communities [15, 127]. Our systematic review provides a first look at research that attempts to make AI systems explainable to communities and stakeholders in the Global South. In this section, we synthesize our findings to discuss (1) the current state of XAI research in the Global South, (2) data and model considerations for building XAI systems for the Global South, and (3) the need for human-centered approaches to XAI in the Global South.

5.1 The state of XAI in the Global South

Our findings show a number of positive trends across existing research that has focused on XAI in the Global South. First, much of the research involves authors from institutions based in this region. This is encouraging since researchers who are from, or have spent significant time in the domains where technologies will be deployed are more likely to understand local contexts. Second, the researchers generally took steps to ensure that their work involved (at least some) data that was local to the context, rather than assuming that data from the Global North would be applicable to contexts in the Global South. In addition, most of the current research focused on advancing social progress by focusing on critical domains, such as healthcare, education, and economic mobility, and targeted a very diverse range of potential users and communities.

However, our findings also highlight concerning gaps that exist in this nascent body of work. First, the small amount of research yielded by our review itself is concerning; we found only 16 papers on XAI in the Global South compared to thousands that study XAI in the Global North [88, 97, 117, 119, 120], highlighting the urgent need to advance the state of XAI in the Global South. Second, we saw a lack of empirical studies conducted with local communities who would be the stakeholders and primary users of XAI systems. Failing to understand the nuances that affect real-world XAI implementations could exacerbate existing inequities, a trend that has already been seen in the deployment of AI systems in the Global North [22, 107, 122]. Additionally, there is little understanding of how technology users in the Global South perceive and interact with XAI technologies. This lack of research leaves the field with significant knowledge gaps as AI practitioners may not have the background knowledge on how to design and develop XAI systems for users in low-resource contexts. We thus call for researchers to engage communities throughout the project lifecycle and conduct studies with the intended users and populations who will be directly impacted by implementation of such systems. Although the same critique regarding the lack of engagement with end users might be true of much AI research, it is especially important in the contexts we are concerned with, where assumptions about prior experience with technology/AI may not hold, and where frontline workers may be expected to be able to critically use these technologies in their work in high-stakes domains.

However, current practices within the peer review process for AI and ML conferences and journals do not incentivize practical implementations of intelligent systems, often favoring advances in benchmarks over real-world deployments [18, 72, 146]. As research within XAI continues to expand into the Global South, these unbalanced incentives may be replicated, preventing researchers from building effective XAI systems in a region where AI deployments are already not understood well. Increasing scientific investments to support advances in XAI research in the Global South as well as realigning incentives to encourage replication and practical implementation could improve how XAI is used within this region and enable researchers to build effective, context-specific tools.

There is also considerable expense to testing XAI systems with users in real-world settings and such work can be limited to institutions that have both the human and physical infrastructure to pursue deployments. For example, the one paper in our corpus that puts an XAI system into practice [125] was spearheaded by IBM India, an institution that has a wealth of resources and is a global leader in AI research. Given the additional complexity that XAI techniques can add to AI models, having the necessary financial and infrastructural resources to run computationally heavy models is an expense that some institutions within the Global South may not be able to bear. Along with increased efforts from HCI and AI practitioners to actively conduct XAI-centered research in the Global South, it will take considerable effort from the broader AI/ML community to embrace the value of making XAI practical for real-world use in the Global South, and throughout other low-resource domains.

5.2 Data and model considerations for building explainable AI for the Global South

Our analysis suggests a need for AI and HCI practitioners to take into account certain considerations when designing and building explainable systems. These considerations center on three things: data, models, and users. In terms of the data that is needed to create XAI systems, as mentioned previously, all of the papers in our review that involved technical implementations of XAI methods

made an effort to use at least some data that was local to the context. However, several of these papers discussed the challenges, additional labor, and associated burdens of needing to also create the datasets that made their work possible. As one example, Sajja et al. [125] detail the additional work needed to build XAI systems of comparable quality to baseline AI models.

The lack of data relevant to local contexts, especially in the Global South, could hamper effective deployments of XAI systems. Using data that is pertinent to local contexts to train ML models and build XAI systems ensures that problems that such tools are aiming to address do so effectively. However, as several papers in our review discussed, getting access to such data or being able to build context-relevant datasets is often not an easy task. Combined with the data-intensive nature of XAI methods, the limitations that exist for practitioners in low-resource regions when collecting, processing, and integrating data into ML pipelines will need to be carefully accounted for in any deployments of XAI systems in the Global South. A growing area of scholarship has begun to focus on understanding the factors that impact data practices in the Global South [1, 62, 114, 127, 142], providing a more empirical view into the issues of scarce and low quality data experienced by researchers in low-resource regions. In their work interviewing AI practitioners in India, East and West African countries, and the United States, Sambasivan et al. [127] push for a move towards "data excellence" where HCI methods are incorporated early and throughout AI data practices and partnerships are formed between AI practitioners, application-domain experts, and field partners to prioritize data collection, documentation, and maintenance. In line with the suggestions above, we propose that increasing training in data literacy for the target end users of AI systems in the Global South and aligning the use of empirical methods in HCI specifically with the goals of XAI can make progress possible in ameliorating existing data challenges.

We also see a need for AI developers and researchers to focus on building new methods of explainability that account for infrastructural and model limitations. Recent failures in the integration of AI systems in the Global South [15] highlight issues that plague AI deployments in these regions, such as the lack of understanding of local contexts before attempting deployment, or incorporating training data that does not reflect real-world input data. These issues could also impact XAI, leading to XAI systems that provide unintelligible explanations or that are structurally infeasible to run in low-resource environments. Of course, this does not mean that researchers should refrain from pursuing XAI research in the Global South—only that there is plenty of room for novel methods to emerge that take into account the constraints presented by these contexts.

Finally, the small number of papers in our review that engaged with target communities point to additional data and model considerations required when building XAI for the Global South. For example, Okolo et al. [108] discuss potentially problematic power dynamics that might arise when deploying AI systems, suggesting that people could potentially defer to decisions made by intelligent systems and assume that the system, rather than their own judgment, is correct. These considerations highlight a need for work that seeks to understand the intricacies of presenting information from XAI systems in ways that balance power dynamics while enabling users to retain their autonomy in choosing whether to follow automated predictions. Of course, XAI can only be truly useful and accessible if it meets the needs of end users, particularly those that do not fit the traditional archetype of an AI practitioner or ML developer. This suggests a need for more human-centered approaches to XAI in the Global South, as we now discuss.

5.3 Towards human-centered explainable AI in the Global South

Our findings and discussion indicate the need for an overhaul of current practices within XAI development. While some work in XAI has begun to explore human-centric approaches to providing explanations via intelligent interfaces [45, 99, 133, 139**?**], the lack of such papers in our review suggests that the vast majority of this work is situated in the Global North and focuses on users who are relatively technology-literate. Technology literacy takes many shapes [89, 90, 95] but is a factor that impacts the adoption of technology [153] and is likely to play a significant role in the adoption of XAI systems and tools.

Of course, it is important to note that the audience of XAI will vary depending on whether the explanations are targeted towards the developers of these systems or human end users (some of whom may have limited experience interacting with technology) and ensuring that their respective needs are addressed. We see rich opportunities for future human-centered XAI research that specifically explores how to explain AI decisions to people with low levels of technology literacy, while also ensuring that these techniques are computationally feasible in low-resource regions. Researchers will further need to engage with questions and/or explore novel methods surrounding how to explain model outputs (e.g., precision, accuracy, ROC, AUC, etc.) and model decisions in languages and formats accessible to low-income, low-literate users who are increasingly seen as a target for AI/ML interventions for Social Good. We acknowledge that difficulty understanding explanations from XAI systems is not limited solely to low-literate users within the Global South. Currently, many state-of-the-art XAI systems provide explanations that only people with machine learning expertise are able to understand. As approaches within the field continue to move toward making AI explainable for users with varying levels of technical knowledge [27, 37, 48, 55, 99, 138**?** , 139], we hope that XAI will no longer remain a grand challenge.

While the potential harms of AI to marginalized populations have been discussed thoroughly in literature [19, 22, 39, 104], little work has been done to understand the specific harms of XAI to marginalized communities. Work by Ehsan and Riedl [36] introduces the concept of "explainability pitfalls" where AI explanations may unintentionally cause users to defer to decisions from AI systems, disregarding their own judgment. In the Global South, explainability pitfalls could particularly affect users who have little technical knowledge and experience operating AI technologies. Prior work has illustrated this in the context of a proposed AI system to diagnose pneumonia [108]. If an explanation for an incorrect diagnosis from an AI system is presented to a community health worker with low AI knowledge, there is a possibility that they would be likely to accept such a decision due to misplaced trust and over-estimation of AI capabilities.

Making AI explainable can be a viable pathway for making AI more useful in real-world environments and increasing its respective ability to improve pressing social issues in domains such as agriculture, healthcare, and education. Knowing the circumstances that affect how different groups of people across various settings understand model decision-making will aid in making XAI systems more responsive to their respective needs. Over the next few decades, there will be tens of thousands of works published in the field of XAI. We envision a future of XAI in the Global South that centers humans while acknowledging the specific constraints associated with deploying such systems in regions where access to data, limited computing infrastructures, and low AI literacy are present. Engaging with communities most likely to benefit from XAI through fieldwork and ethnography and then using such insights to train AI practitioners on the needs of these users could help align the values of local communities with those of the systems that aim to serve them. While it will take significant effort and collaboration to reach this goal, it is critical that the broader AI and HCI research communities converge on a research agenda that will address existing inequities and actively work to solve them in inclusive, sustainable ways.

6 CONCLUSION

This paper contributes a systematic review of emerging XAI research focused on contexts in the Global South. Our search methodology identified 16 papers that, together, highlight both encouraging trends and concerning gaps emerging in XAI research in the Global South. We present a detailed look at these 16 papers, including when and where the research was conducted, datasets used, research approaches taken, communities targeted, and whether the work was put into practice or not. We then discuss considerations for expanding XAI research in low-resource settings and with communities who may have low levels of technology literacy. Our findings suggest new directions for XAI research to ensure that the field addresses growing inequities and makes AI explainable to people from diverse, global contexts.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation grant #1748903, as well as a Google Faculty Research Award. We would also like to thank Priya Pradhan for her work on an alternate version of this paper.

REFERENCES

[1] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. 2021. Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 329–341.

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

[3] Ajay Agarwal and Basant Agarwal. 2021. Tracking Peaceful Tractors on Social Media–XAI-enabled analysis of Red Fort Riots 2021. *arXiv preprint arXiv:2104.13352* (2021).

[4] Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. 2017. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 645–654.

[5] Emily Aiken, Suzanne Bellue, Dean Karlan, Christopher R Udry, and Joshua Blumenstock. 2021. *Machine learning and mobile phone data can improve the targeting of humanitarian assistance*. Technical Report. National Bureau of Economic Research.

[6] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. 2019. iNNvestigate neural networks! *J. Mach. Learn. Res.* 20, 93 (2019), 1–8.

[7] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI. *arXiv preprint arXiv:2106.07483* (2021).

[8] Amazon. 2020. Model Explainability - Amazon SageMaker. Retrieved October 19, 2021 from https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html

[9] Yaw Anokwa, Thomas N Smyth, Divya Ramachandran, Jahanzeb Sherwani, Yael Schwartzman, Rowena Luk, Melissa Ho, Neema Moraveji, and Brian DeRenzi. 2009. Stories from the field: Reflections on HCI4D experiences. *Information Technologies & International Development* 5, 4 (2009), pp–101.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[11] Chinmayi Arun. 2019. AI and the Global South: Designing for Other Worlds. (2019).

[12] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilovic, et al. 2020. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *J. Mach. Learn. Res.* 21, 130 (2020), 1–6.

[13] Raghav Awasthi, Prachi Patel, Vineet Joshi, Shama Karkal, and Tavpritesh Sethi. 2020. Learning explainable interventions to mitigate hiv transmission in sex workers across five states in india. *arXiv preprint arXiv:2012.01930* (2020).

[14] Jeffrey Bardzell and Shaowen Bardzell. 2015. Humanistic hci. *Synthesis Lectures on Human-Centered Informatics* 8, 4 (2015), 1–185.

[15] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[16] Valentina Bellemo, Zhan W Lim, Gilbert Lim, Quang D Nguyen, Yuchen Xie, Michelle YT Yip, Haslina Hamzah, Jinyi Ho, Xin Q Lee, Wynne Hsu, et al. 2019. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health* 1, 1 (2019), e35–e44.

[17] Carlos Eduardo Beluzo, Everton Silva, Luciana Correia Alves, Rodrigo Campos Bresan, Natália Martins Arruda, Ricardo Sovat, and Tiago Carvalho. 2020. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors. *Informatics in medicine unlocked* 20 (2020), 100398.

[18] Yoshua Bengio. 2020. Time to rethink the publication process in machine learning. Retrieved January 21, 2021 from https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/

[19] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces* (2019).

[20] Przemysław Biecek. 2018. DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research* 19, 1 (2018), 3245–3249.

[21] Emma Brunskill, Sunil Garg, Clint Tseng, Joyojeet Pal, and Leah Findlater. 2010. Evaluating an adaptive multi-user educational tool for low-resource environments. In *Proceedings of the IEEE/ACM International Conference on Information and Communication Technologies and Development*. 13–16.

[22] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[23] Jenna Burrell and Kentaro Toyama. 2009. What constitutes good ICTD research? *Information Technologies & International Development* 5, 3 (2009), pp–82.

[24] Ruth MJ Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.. In *IJCAI*. 6276–6282.

[25] William Caicedo-Torres, Ángel Paternina, and Hernando Pinzón. 2016. Machine learning models for early dengue severity prediction. In *Ibero-American Conference on Artificial Intelligence*. Springer, 247–258.

[26] Mary Carman and Benjamin Rosman. 2021. Applying a principle of explicability to AI research in Africa: should we do it? *Ethics and Information Technology* 23, 2 (2021), 107–117.

[27] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[28] Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible AI algorithms: issues, purposes, and challenges. *Journal of Artificial Intelligence Research* 71 (2021), 1137–1181.

[29] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. 2021. A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 1 (2021), e1391.

[30] Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence* 3, 2 (2021), 111–115.

[31] Nicola Dell and Neha Kumar. 2016. The ins and outs of HCI for development. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2220–2232.

[32] Skater developers and contributors. 2017. Skater. Retrieved October 19, 2021 from https://github.com/oracle/Skater

[33] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623* (2018).

[34] Charles-Olivier Dufresne-Camaro, Fanny Chevalier, and Syed Ishtiaque Ahmed. 2020. Computer vision applications and their ethical risks in the global south. (2020).

[35] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[36] Upol Ehsan and Mark O Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[37] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[38] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[39] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[40] Juliana Jansen Ferreira and Mateus de Souza Monteiro. 2020. Do ML Experts Discuss Explainability for AI Systems? A discussion case in the industry for a domain-specific solution. *arXiv preprint arXiv:2002.12450* (2020).

[41] David Barrera Ferro, Sally Brailsford, Cristián Bravo, and Honora Smith. 2020. Improving healthcare access management by predicting patient no-show behaviour. *Decision Support Systems* 138 (2020), 113398.

[42] The Institute for Ethical AI & ML. 2021. XAI - An eXplainability toolbox for machine learning. Retrieved October 21, 2021 from https://github.com/EthicalML/xai

[43] Rob Geada, Tommaso Teofili, Rui Vieira, Rebecca Whitworth, and Daniele Zonca. 2021. TrustyAI Explainability Toolkit. *arXiv preprint arXiv:2104.12717* (2021).

[44] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. 2022. Explainable AI, But Explainable to Whom? An Exploratory Case Study of xAI in Healthcare. In *Handbook of Artificial Intelligence in Healthcare*. Springer, 169–198.

[45] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[46] Indranil Ghosh and Manas K Sanyal. 2021. Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights* (2021), 100039.

[47] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[48] Google. 2017. AI Experiments - Experiments with Google. Retrieved October 19, 2021 from https://experiments.withgoogle.com/collection/ai

[49] Google. 2020. Explainable AI | Google Cloud. Retrieved October 19, 2021 from https://cloud.google.com/explainable-ai

[50] Google. 2021. Responsible AI Toolkit | TensorFlow. Retrieved October 19, 2021 from https://www.tensorflow.org/responsible_ai

[51] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[52] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[53] Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, and Jeremy De Beer. 2020. Artificial intelligence (AI) deployments in Africa: benefits, challenges and policy dimensions. *The African Journal of Information and Communication* 26 (2020), 1–28.

[54] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2970–2979.

[55] Yen-Chia Hsu, Ting-Hao'Kenneth' Huang, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, and Alessandro Bozzon. 2021. Empowering Local Communities Using Artificial Intelligence. *arXiv preprint arXiv:2110.02007* (2021).

[56] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. 2021. Model complexity of deep learning: A survey. *Knowledge and Information Systems* 63, 10 (2021), 2585–2619.

[57] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[58] IBM. 2020. Explainable AI. Retrieved March 31, 2021 from https://www.ibm.com/watson/explainable-ai

[59] Deep Learning Indaba. 2021. Deep Learning Indaba. Retrieved October 21, 2021 from https://deeplearningindaba.com/2021/

[60] Lucas Introna and David Wood. 2004. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society* 2, 2/3 (2004), 177–198.

[61] Tupokigwe Isagah and Albogast Musabila. 2020. Recommendations for artificial intelligence implementation in African governments: results from researchers and practitioners of AI/ML. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*. 82–89.

[62] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[63] Azra Ismail and Neha Kumar. 2021. AI in Global Health: The View from the Front Lines. (2021).

[64] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive Explanations for Model Interpretability. *arXiv preprint arXiv:2103.01378* (2021).

[65] Bhanu Jain, Manfred Huber, Leonidas Fegaras, and Ramez A Elmasri. 2019. Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 599–607.

[66] Kuhu Joshi and Chaitanya K Joshi. 2019. Working women and caste in India: A study of social disadvantage using feature attribution. *arXiv preprint arXiv:1905.03092* (2019).

[67] Hyungsik Jung and Youngrock Oh. 2021. Towards Better Explanations of Class Activation Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1336–1344.

[68] Ashish Kapoor, Nathan Eagle, and Eric Horvitz. 2010. People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In *2010 AAAI Spring Symposium Series*.

[69] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 45–55.

[70] Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *International Conference on Case-Based Reasoning*. Springer, 163–178.

[71] Niklas Keller, Mirjam A Jenny, Claudia A Spies, and Stefan M Herzog. 2020. Augmenting Decision Competence in Healthcare Using AI-based Cognitive Models. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–4.

[72] Hannah Kerner. 2020. Too many AI researchers think real-world problems are not relevant. Retrieved January 21, 2022 from https://www.technologyreview.com/2020/08/18/1007196/ai-research-machine-learning-applications-problems-opinion/

[73] Os Keyes, Josephine Hoy, and Margaret Drouhard. 2019. Human-computer insurrection: Notes on an anarchist HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[74] Masoud Bakhtyari Kia, Saied Pirasteh, Biswajeet Pradhan, Ahmad Rodzi Mahmud, Wan Nor Azmin Sulaiman, and Abbas Moradi. 2012. An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. *Environmental earth sciences* 67, 1 (2012), 251–264.

[75] Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. 2021. Alibi Explain: algorithms for explaining machine learning models. *Journal of Machine Learning Research* 22, 181 (2021), 1–7.

[76] Akhil Alfons Kodiyan. 2019. An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint* (2019).

[77] Tino Kreutzer, Patrick Vinck, Phuong N Pham, Aijun An, Lora Appel, Eric DeLuca, Grace Tang, Muath Alzghool, Kusum Hachhethu, Bobi Morris, et al. 2019. Improving humanitarian needs assessments through natural language processing. *IBM Journal of Research and Development* 64, 1/2 (2019), 9–1.

[78] Nicolas Lainez. 2021. The Prospects and Dangers of Algorithmic Credit Scoring in Vietnam: Regulating a Legal Blindspot. (2021).

[79] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

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

[81] Chiara Ledesma, Oshean Lee Garonita, Lorenzo Jaime Flores, Isabelle Tingzon, and Danielle Dalisay. 2020. Interpretable Poverty Mapping using Social Media Data, Satellite Images, and Geospatial Information. *arXiv preprint arXiv:2011.13563* (2020).

[82] Kwang Hee Lee, Chaewon Park, Junghyun Oh, and Nojun Kwak. 2021. LFI-CAM: Learning Feature Importance for Better Visual Explanation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1355–1363.

[83] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[84] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[85] Xuan Liu, Xiaoguang Wang, and Stan Matwin. 2018. Improving the interpretability of deep neural networks with knowledge distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 905–912.

[86] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. 2020. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 1–16.

[87] Loyani K Loyani, Karen Bradshaw, and Dina Machuve. 2021. Segmentation of Tuta Absoluta's Damage on Tomato Plants: A Computer Vision Approach. *Applied Artificial Intelligence* 35, 14 (2021), 1107–1127.

[88] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[89] Hughie Mackay. 1992. From computer literacy to technology literacy. *Technological literacy and the curriculum* (1992), 125–147.

[90] Guy Merchant. 2007. Writing the future in the digital age. *Literacy* 41, 3 (2007), 118–128.

[91] Mvurya Mgala and Audrey Mbogho. 2015. Data-driven intervention-level prediction modeling for academic performance. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*. 1–8.

[92] Microsoft. 2019. Model interpretability in Azure Machine Learning. Retrieved October 19, 2021 from https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability#interpretability-with-azure-machine-learning

[93] Mona Minakshi, Pratool Bharti, Willie B McClinton III, Jamshidbek Mirzakhalov, Ryan M Carney, and Sriram Chellappan. 2020. Automating the surveillance of mosquito vectors from trapped specimens using computer vision techniques. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 105–115.

[94] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. 2021. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law* 29, 2 (2021), 111–147.

[95] David Richard Moore. 2011. Technology literacy: The extension of cognition. *International Journal of Technology and Design Education* 21, 2 (2011), 185–193.

[96] Mário WL Moreira, Joel JPC Rodrigues, Francisco HC Carvalho, Naveen Chilamkurti, Jalal Al-Muhtadi, and Victor Denisov. 2019. Biomedical data analytics in mobile-health environments for high-risk pregnancy outcome prediction. *Journal of Ambient Intelligence and Humanized Computing* 10, 10 (2019), 4121–4134.

[97] Satya M Muddamsetty, NS Jahromi Mohammad, and Thomas B Moeslund. 2020. Sidu: Similarity difference and uniqueness method for explainable ai. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3269–3273.

[98] R Muhamedyev, K Yakunin, YA Kuchin, A Symagulov, T Buldybayev, S Murzakhmetov, and A Abdurazakov. 2020. The use of machine learning "black boxes" explanation systems to improve the quality of school education. *Cogent Engineering* 7, 1 (2020), 1769349.

[99] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. 2019. Efficient saliency maps for Explainable AI. *arXiv preprint arXiv:1911.11293* (2019).

[100] Charles K Mutai, Patrick E McSharry, Innocent Ngaruye, and Edouard Musabanganji. 2021. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC medical research methodology* 21, 1 (2021), 1–11.

[101] Mobeen Nazar, Muhammad Mansoor Alam, Eiad Yafi, and MS Mazliham. 2021. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. *IEEE Access* (2021).

[102] Ye Ni, Xutao Li, Yunming Ye, Yan Li, Chunshan Li, and Dianhui Chu. 2020. An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction. *IEEE Geoscience and Remote Sensing Letters* (2020).

[103] Data Science Nigeria. 2019. Annual Report July 2018–June 2019: Building One Million AI Talents in 10 Years. Retrieved October 21, 2021 from https://www.datasciencenigeria.org/wp-content/uploads/2019/08/annual-report-final.pdf

[104] Safiya Umoja Noble. 2018. *Algorithms of oppression*. New York University Press.

[105] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[106] Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A Bakker, Luis Tejerina, and Alex Pentland. 2020. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 241–251.

[107] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[108] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. " It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. (2021).

[109] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–Machine Translation For Africa. *arXiv preprint arXiv:2003.11529* (2020).

[110] Shashwat Pathak and Basant Kumar. 2016. A robust automated cataract detection algorithm using diagnostic opinion based parameter thresholding for telemedicine application. *Electronics* 5, 3 (2016), 57.

[111] Rabin Patra, Joyojeet Pal, and Sergiu Nedevschi. 2009. ICTD state of the union: Where have we reached and where are we headed. In *2009 International Conference on Information and Communication Technologies and Development (ICTD)*. IEEE, 357–366.

[112] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[113] Chris Percy, Simo Dragicevic, Sanjoy Sarkar, and Artur S Garcez. 2021. Accountability in AI: From Principles to Industry-specific Accreditation. *arXiv preprint arXiv:2110.09232* (2021).

[114] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. 2019. Examining the challenges in development data pipeline. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 13–21.

[115] Thomas Ploug and Søren Holm. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine* 107 (2020), 101901.

[116] Alex Polacco and Kayla Backes. 2018. The amazon go concept: Implications, applications, and sustainability. *Journal of Business and Management* 24, 1 (2018), 79–92.

[117] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[118] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[119] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[120] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[121] Eugenia Arrieta Rodríguez, Francisco Edna Estrada, William Caicedo Torres, and Juan Carlos Martínez Santos. 2016. Early prediction of severe maternal morbidity using machine learning techniques. In *Ibero-American Conference on Artificial Intelligence*. Springer, 259–270.

[122] Casey Ross. 2022. Epic's AI algorithms, shielded from scrutiny by a corporate firewall, are delivering inaccurate information on seriously ill patients. *STAT* (2022). https://www.statnews.com/2021/07/26/epic-hospital-algorithms-sepsis-investigation/

[123] Mark Ryan, Josephina Antoniou, Laurence Brooks, Tilimbe Jiya, Kevin Macnish, and Bernd Stahl. 2020. The ethical balance of using smart information systems for promoting the United Nations' Sustainable Development Goals. *Sustainability* 12, 12 (2020), 4826.

[124] Mazrura Sahani and Zainudin Mohd Ali. 2017. Feature selection algorithms for Malaysian dengue outbreak detection model. *Sains Malaysiana* 46, 2 (2017), 255–265.

[125] Shravan Sajja, Nupur Aggarwal, Sumanta Mukherjee, Kushagra Manglik, Satyam Dwivedi, and Vikas Raykar. 2021. Explainable AI based Interventions for Pre-season Decision Making in Fashion Retail. In *8th ACM IKDD CODS and 26th COMAD*. 281–289.

[126] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.

[127] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[128] Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 5–22.

[129] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[130] Sophia L Sanga, Dina Machuve, and Kennedy Jomanga. 2020. Mobile-based deep learning models for Banana disease detection. *Engineering, Technology & Applied Science Research* 10, 3 (2020), 5674–5677.

[131] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[132] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[133] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[134] Nathan Silberman, Kristy Ahrlich, Rob Fergus, and Lakshminarayanan Subramanian. 2010. Case for automated detection of diabetic retinopathy. In *2010 AAAI Spring Symposium Series*.

[135] Moninder Singh, Gevorg Ghalachyan, Kush R Varshney, and Reginald E Bryant. 2021. An Empirical Study of Accuracy, Fairness, Explainability, Distributional Robustness, and Adversarial Robustness. *arXiv preprint arXiv:2109.14653* (2021).

[136] Daniel Smilkov and Shan Carter. 2017. Tensorflow-Neural Network Playground. Retrieved October 19, 2021 from https://playground.tensorflow.org/

[137] Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen. 2020. Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1241–1248.

[138] Yunjia Sun. 2016. *Novice-Centric Visualizations for Machine Learning*. Master's thesis. University of Waterloo.

[139] Yunjia Sun, Edward Lank, and Michael Terry. 2017. Label-and-Learn: Visualizing the Likelihood of Machine Learning Classifier's Success During Data Labeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 523–534.

[140] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. 2018. Learning global additive explanations for neural nets using model distillation. (2018).

[141] Mahyat Shafapour Tehrany, Biswajeet Pradhan, and Mustafa Neamah Jebur. 2014. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of hydrology* 512 (2014), 332–343.

[142] Divy Hasmukhbhai Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is ML data good?: Valuing in Public Health Datafication. (2022).

[143] Dieu Tien Bui, Biswajeet Pradhan, Owe Lofman, and Inge Revhaug. 2012. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and Naive Bayes Models. *Mathematical problems in Engineering* 2012 (2012).

[144] Aditya Vashistha, Richard Anderson, and Shrirang Mare. 2018. Examining security and privacy research in developing regions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–14.

[145] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 1–10.

[146] Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).

[147] Kutoma Wakunuma, Tilimbe Jiya, and Suleiman Aliyu. 2020. Socio-ethical implications of using AI in accelerating SDG3 in Least Developed Countries. *Journal of Responsible Technology* 4 (2020), 100006.

[148] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[149] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[150] Josephine Yam and Joshua August Skorburg. 2021. From human resources to human rights: Impact assessments for hiring algorithms. *Ethics and Information Technology* (2021), 1–13.

[151] Cheryl Young, Stephen Barker, Rodney Ehrlich, Barry Kistnasamy, and Annalee Yassi. 2020. Computer-aided detection for tuberculosis and silicosis in chest radiographs of gold miners of South Africa. *The International Journal of Tuberculosis and Lung Disease* 24, 4 (2020), 444–451.

[152] Meg Young, Michael Katell, and PM Krafft. 2019. Municipal surveillance regulation and algorithmic accountability. *Big Data & Society* 6, 2 (2019), 2053951719868492.

[153] Tai-Kuei Yu, Mei-Lan Lin, and Ying-Kai Liao. 2017. Understanding factors influencing information communication technology adoption behavior: The moderators of information literacy and digital skills. *Computers in Human Behavior* 71 (2017), 196–208.

[154] Rida Zainab and Rajarathnam Chandramouli. 2020. Detecting and Explaining Depression in Social Media Text with Machine Learning. (2020).

[155] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.