From theoretical models to practical deployment: A perspective and case study of opportunities and challenges in AI-driven healthcare research for low-income settings

Felix Krones[1*], Benjamin Walker[2]

1 Oxford Internet Institute, University of Oxford, Oxford, UK
2 Mathematical Institute, University of Oxford, Oxford, UK

*felix.krones@oii.ox.ac.uk

Abstract

This paper critically explores the opportunities and challenges of deploying Artificial Intelligence (AI) in healthcare. This study has two parallel components:

(1) A narrative literature summary, which assesses the capacity of AI to aid in addressing the observed disparity in healthcare between high- and low-income countries. Despite the development of machine learning models for a wide range of diseases, many are never deployed in practice. We highlight various challenges that contribute to the lack of deployed models. A main challenge that is not always sufficiently addressed in the literature is the evaluation of model generalisation. For example, by using a multi-site set-up with test sets that were collected separately to the train and validation sets, or by using evaluation metrics which are both understandable and clinically applicable. Moreover, we discuss how the emerging trend of human-centred deployment research is a promising avenue for overcoming barriers towards deployment.

(2) A case study on developing and evaluating a predictive AI model tailored for low-income environments. The focus of this case study is heart murmur detection in rural Brazil. Our Binary Bayesian ResNet model leverages overlapping log mel spectrograms of patient heart sound recordings and integrates demographic data and signal features via XGBoost to optimise performance. We discuss the model's limitations, its robustness, and the obstacles preventing its practical application. We especially highlight how our model, and other state-of-the-art models, struggle to generalise to out-of-distribution data.

The research accentuates the transformative potential of AI-enabled healthcare, particularly affordable point-of-care monitoring systems, in low-income settings. It also emphasises the necessity for effective implementation and integration strategies to guarantee the successful deployment of these technologies.

Author summary

In this study, we explore the potential and limitations of Artificial Intelligence (AI) in healthcare, focusing on its role in addressing global health inequities.
Non-communicable diseases, especially cardiovascular disorders, are a leading global cause of death, exacerbated in low-income settings due to restricted healthcare access. Our research has two components: a narrative literature summary that discusses the gap between AI research and real-world applications, and a case study on heart murmur detection in rural Brazil. The case study introduces an AI model tailored for

low-income environments, which efficiently analyses heart sound recordings for diagnostic insights. Both parts especially highlight the challenges of generalisation to out-of-distribution data.

Our findings accentuate AI's capability to revolutionise point-of-care monitoring in resource-limited settings. However, they also highlight the critical importance of effective implementation and conscientious design for the successful deployment of these technologies. Through this work, we contribute to the broader objective of fostering global health equity by leveraging AI, while emphasising the need for thoughtful application and integration strategies.

1 Introduction

This paper begins with an introduction to cardiovascular diseases and the PhysioNet Challenge 2022, which forms the basis of our case study. This is followed by an overview of related work and our contributions. Section 2, as part of an extended introduction, offers a narrative literature overview of the opportunities and challenges of AI in healthcare, disparities between income settings, and deployment considerations.

1.1 Cardiovascular diseases

Non-communicable diseases are the leading cause of global mortality. In 2019, 55 million people died; 74% of these deaths were from non-communicable diseases, as opposed to 18% from communicable diseases and 8% from injuries [1]. Cardiovascular diseases, as part of non-communicable diseases, accounted for 17.9 million (or 32%) of global deaths [1, 2]. These figures are more alarming in low- and middle-income countries, where over three-quarters of these deaths occur due to limited access to early detection and effective treatment measures [1–3].

Cardiovascular diseases comprise various heart and vessel disorders, such as coronary artery disease, valvular heart disease, and congenital heart disease [4]. Although coronary artery disease is more common in developed nations, congenital and valvular heart diseases are more prevalent in developing countries due to limited prenatal screening and healthcare access. Annually, rheumatic heart diseases account for over 68 million cases and approximately 1.4 million deaths, primarily affecting children and young adults [4]. Early identification of these diseases is vital as lifestyle changes can prevent a substantial number of cases. However, the lack of robust primary healthcare often leads to late detection and premature deaths.

The World Health Organisation (WHO) aims to reduce the probability of dying from non-communicable diseases between ages 30-69 to 12.3% by 2030, down from 17.8% in 2019 [1]. WHO's strategies include risk factor reduction and improved disease detection. To this end, the WHO has set international objectives, such as lowering the incidence of elevated blood pressure and ensuring 80% availability of affordable basic technologies and medicines for cardiovascular diseases [2]. Achieving these goals necessitates significant investments in health systems, especially in low- to medium-income countries. Thus, cost-effective point-of-care technologies are crucial for heart disease screening in these settings. Encouragingly, initial results indicate a 27% decline in the individual risk for cardiovascular diseases from 2000-2019 [1].

Anomalies in the early stages of heart structure development can lead to congenital heart disease. While most murmurs are innocent, detecting heart murmurs may serve as indicators of these structural defects. Early-life heart sound signal analysis could act as a rapid, non-invasive screening method for cardiac structural anomalies, facilitating prompt diagnosis and treatment [5]. Cardiac auscultation and phonocardiography analysis offer straightforward methods for diagnosing heart conditions by identifying

abnormal sound waves and heart murmurs in heart sound recordings [5]. The initial stages of heart murmur screening can be relatively straightforward with proper guidance [4]. Nurses can effectively be trained to use a stethoscope and record heart sounds with technological support. However, interpreting these sounds requires professionals with years of experience, who may not always be readily available. In such scenarios, AI-assisted pre-screening could serve as a viable solution, aiding in the referral of patients to specialised treatment facilities.

1.2 PhysioNet Challenge context

In the context of the PhysioNet Challenge 2022 [6], our research builds upon the submission of our previous team, PathToMyHeart [7]. However, we have made several adjustments to better align with the broader discourse on deploying point-of-care devices in resource-constrained environments and to enhance our models further.

The objective of the Challenge was to "identify the presence, absence, or unclear cases of murmurs and the normal vs. abnormal clinical outcomes from heart sound recordings" [6]. The scoring mechanism employed a weighted accuracy for the three-class murmur categorisation and a cost function for outcomes classification. The three categories of murmur were: patients where murmurs were present, patients where the presence of murmurs was uncertain, and patients where murmurs were absent, which were assigned weights 5, 3, and 1 in the weighted accuracy respectively. The cost function for the outcome classification incorporated: (a) the expert capacity factor in determining the costs associated with patient screening (i.e., when classified as abnormal), (b) significant costs if a patient exhibited abnormal heart sounds but did not receive treatment, and (c) additional costs for treating a patient [6].

In this paper, we aim to broaden the discourse by focusing on more fundamental error scores that are straightforward to interpret. However, we acknowledge that any real-world deployment would require consideration of cost implications similar to those delineated in the Challenge. We adopt a conservative strategy, treating both problems as binary classification tasks, where unclear murmur cases are considered present. The rationale behind this approach is that the algorithm would primarily serve as a prescreening tool, with healthcare professionals ultimately making the final diagnosis. The most significant limitation of our work, to be discussed later, is that without a known deployment setting, identifying the most appropriate cost function to optimise the models becomes challenging.

1.3 Related work

Recent reviews have revealed that most current approaches in the classification of heart sounds primarily focus on a binary problem: categorising heart sounds as either normal or abnormal. This emphasis largely stems from the scarcity of available heart sound data for more nuanced classifications [8]. While many studies report accuracies exceeding 90% for heart sound classification tasks (cf. Section 5.2), depending on the task and dataset [9], recent reviews particularly highlight the need for further development to establish robust methods. In terms of deployment and robustness, the reviews identify several challenges [8, 9]. First, the complex, non-stationary nature of heart sound signals complicates their extraction and analysis. Second, the introduction of noise and interference during the acquisition process exacerbates these challenges. Third, the reviews indicate that existing algorithms exhibit limited capabilities and inconsistent accuracy rates, suggesting that they may not yet be adequately robust for practical, clinical applications. Importantly, these reviews stress the necessity for evaluation using universally standardised databases for a more accurate comparison of algorithmic performance [9].

1.4 Contributions

In this study, we evaluate deployment challenges in healthcare technologies through a narrative literature overview and a case study. We tailor and assess a predictive AI model for heart murmur detection, focusing on resource-constrained environments in rural areas of low-income countries. Additionally, we examine the model's real-world limitations, evaluate its robustness, and discuss barriers to its practical deployment.

This paper expands upon our previous work [7], which focused on heart murmur classification and received recognition in the 2022 George B. Moody PhysioNet Challenge [6], securing fourth place. The ultimate objective remains the same: to create an open-source algorithm for accurate classification of heart murmurs using heart sound recordings.

In alignment with our overarching research question—*How can AI technologies be effectively deployed to bridge healthcare disparities between high-income and low-income countries, and what are the opportunities and challenges in achieving this goal?*—our contributions are as follows:

- We synthesise findings from a narrative literature overview and our case study. Our focus is on identifying challenges and barriers to deploying AI models for pre-screening in low-income settings. We discuss the findings thoroughly in the discussion Section and assess the challenges using the NASSS framework [10].
- Expanding upon [7], we enhance the deep learning model by incorporating multimodal data to extend its generalisability. We employ two-dimensional spectrograms derived from heart sound recordings for the classification of heart murmurs. We compare the improved model to other architectures, including baseline Residual Networks (ResNets) without a Bayesian component.
- We carry out a multi-site validation of the refined model, perform a robustness evaluation, and critically identify gaps that require attention for successful real-world deployment.

2 AI deployment in healthcare

2.1 Opportunities and challenges of AI in healthcare

AI in healthcare offers a significant opportunity to enhance various areas, such as diagnostics, treatment planning, and overall patient outcomes, particularly in the context of increasing healthcare demands and costs [11, 12]. The capabilities of AI have expanded across various healthcare applications in recent years. For instance, image reconstruction and analysis in radiology has undergone substantial improvement due to the integration of deep neural networks [13]. AI's assistance to healthcare professionals through computer-aided detection and diagnosis has shown to have the potential to improve efficiency and accuracy [14, 15]. Furthermore, algorithms that identify areas of interest during image screening have proven effective by supporting clinicians, thus enhancing the diagnostic process without supplanting human expertise [14].

However, the incorporation of AI into healthcare encounters numerous obstacles. These include regulatory hurdles, data privacy concerns, data quality issues, ethical considerations, clinical validation, and funding shortfalls [16]. From a technical standpoint, it is important that models are robust, adaptable, and accurately convey their uncertainty [17]. Studies have begun to address these issues in more detail, especially focusing on robustness towards out-of-distribution shifts; for example, by evaluating model performance across various datasets [18], and by examining specific data changes, such as temporal variations [19]. But this also means looking at more

metrics than just AUC [17], especially metrics that are both understandable and clinically applicable.

In recent literature, a special interest is given to population shifts, especially to a model's performance across different patient subgroups, to ensure fairness and address biases that often stem from unbalanced training data [20, 21]. Future research should aim to improve representation of underrepresented groups and rare conditions, thus advancing fairness and efficacy. Attaining improved generalisation is another vital step towards this goal [22]. This involves considering the impact of varying error rates beyond a narrow set of fairness metrics and acknowledging additional factors like absolute welfare or priority. This is important to prevent Pareto-inefficient outcomes, a situation where enhancements in model performance for one group could still be realised without negatively impacting other groups [20].

Besides a more realistic model evaluation and developing approaches for better generalisation [23], addressing the mentioned challenges necessitates a concerted effort that includes education, collaboration among healthcare providers and industry stakeholders, as well as ongoing evaluation and refinement of AI systems [24]. To foster acceptance among end-users, early integration of them into development is crucial, as is training to work effectively with the technology [16, 25]. In addition to meeting performance expectations, it is essential to meet the expectations regarding the required effort, the social impact of the systems (e.g., on communication or decision-making), and other facilitating conditions such as infrastructure and legal frameworks [25].

2.2 Out-of-distribution performance

Model reliability on out-of-distribution data, which was unseen during training, is a huge concern in medical imagining and medical AI research. The data distribution of medical images can change, for example, due to the variations in imaging equipment, the use of different protocols, or changes in the patient populations across locations and time [23].

To address out-of-distribution shifts, multi-site testing is a critical step before deploying AI applications in healthcare. It allows for the assessment of model generalisability, which is essential for any AI tool to work effectively across various locations and populations. This is particularly important in healthcare, where variability in patient demographics, disease prevalence, and clinical practices, including data collection, can significantly impact model performance. Despite its importance, research indicates that a substantial 72% of recent clinical machine learning studies have not conducted multi-site evaluation, suggesting a considerable gap in the current approach to AI deployment in healthcare; many AI models initially appearing to outperform human practitioners failed to maintain their superiority under more rigorous multi-site testing [26]. This underlines the potential risk of over-reliance on single-site evaluations and highlights the necessity for extensive multi-site testing. Moreover, multi-site evaluation plays a pivotal role in identifying and adapting to distribution shifts over time, which is crucial for maintaining the robustness of the model, ensuring that it continues to perform effectively in changing environments.

The gold standard in testing AI models is using randomised controlled trials. However, these trials have only been conducted in a few dozen studies [27]. One study, for example, used a randomised trail to investigate performance, costs, and especially treatment time for HIV-Tuberculosis screening in Malawi [28]. While many studies only involve small cohorts, they mark a significant step towards aligning machine learning research in healthcare with real-world applications.

The challenge in our study resides in testing the heart murmur model across multiple sites using publicly available data. As illustrated in Table 1, not many databases exists which are comparable in size to the 2022 Challenge data. Those that are available do not all contain multimodal data and have often only much shorter and cleaner

Name	Rec. [#]	Freq. [Hz]	Durat. [sec]	Labels	Patients [#]	Location	Demographics
PhysioNet 22 [4]	5272	4000	5 - 80	Murmur: Present Absent Unknown Outcome: Normal Abnormal	1568	Available	Available
PhysioNet 16 [29]	3153	2000	5 - 120	Normal Abnormal	764	Partially	Partially
Yaseen [30]	1000	8000	1 - 4	Normal Aortic Stenosis Mitral Stenosis Mitral Regur. Mitral Prolapse	na	na	na
Pascal B [31]	656	4000	1 - 25	Normal Murmur Extrasystole	na	na	na

Table 1. Overview of publicly available heart sound databases with more than 500 recordings.

recordings available. Given that the PhysioNet/CinC Challenge 2016 database is the largest available resource with multimodal data, we chose it for our multi-site evaluation. As most databases primarily contain labels for normal/abnormal classification, we decided to concentrate on this outcome task for our multi-site assessment.

2.3 Healthcare AI in low-income settings

Despite considerable strides towards achieving the health-related sustainable development goals set by the WHO [1], a pronounced discrepancy still exists between the health outcomes and available health resources in high-income countries (HICs) and their low- or medium-income counterparts (LMICs). For instance, in 2020, a global shortfall of 15 million health workers was reported [1], a gap that is notably wider in LMICs than in HICs; the disparity is stark: Europe reported an average of 36.6 medical doctors per 10,000 citizens, a figure that contrasts sharply with a mere 2.9 in Africa and 7.7 in South-East Asia.

Such disparities highlight the diverse needs and potential applications of AI technologies across different resource settings. In developed nations, a major use case of AI is improving the individualisation an efficiency of healthcare. In contrast, in low-income settings, a major use of AI is serving as a bridge to close healthcare delivery gaps. For example, while citizens in HICs may have immediate access to medical professionals, a pressing need exists for simplified pre-screening systems in LMICs, which can be administered by frontline healthcare workers. AI can facilitate task shifting, enabling community health workers to deliver more services [32]. Technologies like AI-driven heart sound interpretation can offer initial pre-screening for cardiac conditions in areas where doctors are scarce. Consequently, AI has the potential to significantly enhance both the quality and quantity of healthcare in LMICs [33–36].

A myriad of machine learning models for healthcare have been developed recently, with a significant emphasis on aiding LMICs. Examples include COVID-19 forecast models in Iran [37] and India [38], Ebola forecast models for Africa [39], and automated malaria diagnostics in Uganda [40]. Various tuberculosis prediction studies in Brazil [41], South Africa [42] and Peru [43] have been conducted. Furthermore, extensive studies

have been performed on diabetic retinopathy in India [44, 45] and their transferability from HICs to LMICs. For instance, an AI model trained to detect diabetic retinopathy on data collected in Singapore has been shown to maintain its effectiveness when evaluated on data collected in Zambia, demonstrating that well-developed AI can be a valuable resource even across sites [35]. These applications illustrate AI's transformative potential in healthcare, enabling affordable, accessible, and high-quality care.

Nonetheless, deploying AI technologies in LMICs is not without hurdles. There have been reports indicating concerns such as inconsistent reliability, varied effects on operational processes, a deficiency in user-centred design, and an incompatibility with regional particularities [33, 34, 46]. Other barriers to the successful development and adoption of high-performing AI solutions, which are tailored to local settings, encompass constraints in data accessibility and demonstrable financial viability [36, 47], as well as concerns surrounding the openness of the data and computation methods involved in training AI tools [46].

In conclusion, while AI holds significant potential for healthcare in LMICs, considerable challenges exist. It is important to conduct further assessments on the incorporation of AI within the healthcare sector in LMICs. This will help in ascertaining its efficiency and dependability in practical environments and contribute to the development of insights concerning optimal strategies for upcoming deployments.

Since diagnostic decision making is one of the most considered and promising applications for AI in low-income settings [32], we chose it as our focus for this study.

2.4 Emphasising human-centred deployment

While many studies demonstrate impressive performance on paper, a notable gap often exists towards actual deployment. Hence, current trends are leaning towards human-centred deployment research. Research and our experience in implementing data science projects have shown that the early integration of the end-users can foster a wider acceptance of the technology [25, 48, 49]. Notable projects such as Google's automated retinal disease assessment in Thailand and India [49, 50] are garnering significant attention. This project, in collaboration with various clinics, conducted a human-centred observational study to examine the consequences of the algorithm's implementation on clinical processes, and to identify the elements influencing the performance of the system's algorithm. By 2023, it had screened more than 200,000 people, revealing challenges in data quality, workflow integration, and post-deployment monitoring when shifting into the real world.

More research efforts have begun to assess the deployment of AI technologies in LMICs: Studies by [48] and [51] have examined AI usage among frontline health workers in India, highlighting key design considerations for future applications. Other investigations have conducted pilot studies involving frontline healthcare workers for tasks such as triaging palpable breast lumps in Mexico using an AI-based computer-assisted diagnosis with a low-cost portable ultrasound system [52], and automated radiation planning in South Africa to reduce maximum dosage in cervical cancer treatment [53]. Despite the mixed performance of an AI tool in COVID CT diagnosis in Ecuador, it remained in use due to the absence of alternatives [54], an issue to be addressed in future applications to maintain user trust.

2.5 Examples from practice

The objective of many recently developed technologies for LMICs is either to assist frontline healthcare workers [51], e.g., with user-friendly screening tools, or to aid non-specialist clinicians, e.g., non-radiologists, with tools that assist in the analysis of X-rays. For instance, recent work by [55] showed that an AI system for chest radiograph

interpretation, when combined with input from a non-radiology resident, achieved performance metrics comparable to those of board-certified radiologists.

As [48] has noted, many studies are steered by large tech companies, such as the before mentioned study by Google [49, 50]. However, an increasing number of smaller companies are now also attempting to deploy AI technologies.

Several enterprises and organisations are collaborating with researchers to deploy AI technologies in LMICs, such as Wadhwani AI [56], an Indian company focusing on AI solutions to reduce morbidity and mortality among mothers and children, in addition to other eHealth, dermatology, and ophthalmology tools. Aidoc [57], based in Israel/US, is developing AI tools for cardiovascular and neuroscience diseases with a focus on radiology, care coordination, patient management, and clinical trial enrolment. Ubenwa AI [58], a Nigeria/Canada based company, is developing a computer-aided diagnostics for perinatal asphyxia using infant cry sounds. Other organisations, such as OpenMRS [59] and DHIS2 [60], provide medical record systems to countries worldwide.

An example which is directly focused on tackling the hurdles associated with AI implementation for medical imaging in resource-scarce settings is RAD-AID–a non-profit committed to enhancing radiology resources in low-income environments [61]. The tackled hurdles include a shortage of equipment, professional expertise, infrastructure, and defined data-rights policies; moreover, the compromise of the trustworthiness of AI by a lack of data diversity and opaque algorithms. RAD-AID introduces a triad strategy of clinical radiology education, infrastructure establishment, and staggered AI introduction. The organisation highlights that AI implementation in LMIC necessitates a different strategy compared to HIC due to variations in resources and clinical scenarios.

However, with regard to the real-life diagnostic accuracy of commercially available tools more research has to be done. [62] give a small outlook: On the example of four chest radiography AI tools they evaluated the tools' performances on 2,040 patients. Their findings indicate that tools are designed to act rather conservative: While the authors report moderate to high sensitivity, they also noticed a production of more false-positive findings than comparable radiology reports, and decreasing performance for smaller targets and for cases with multiple findings. This highlights the opportunities for AI based screening methods but brings out the need for a careful deployment at the same time.

3 Materials and Methods

3.1 Training data

In this research, we use the data collected by [4]. The heart sound recordings were gathered using a Littmann 3200 stethoscope (a quick online search revealed that this stethoscope is priced between 250 and 300 pounds in the UK [63]) and the tablet-based GUI software, DigiScope Collector. This software provides a user-friendly interface for collecting patient metadata and offers clear guidance on the process of recording heart sounds. Heart sound recordings of 1,568 individuals were obtained from an initial pool of 2,061 participants during two screening campaigns in 2014 and 2015. These campaigns, known as 'Caravana do Coração', took place in the state of Paraíba in northern Brazil. Participants were filtered based on eligibility criteria. Mobile teams travelled across the state during the campaigns, collecting data predominantly from a pediatric population. Notably, 63% of the participants were children, and 20% were infants [4]. From the original 1,568 patients, 53.2% were referred for a follow-up, while 36.7% were discharged entirely. The remaining 10.1% needed additional testing (27 patients), were indicated for surgery or intervention (35 patients), or no information were recorded (97 patients).

The dataset includes heart sound recordings ranging from 5 to 80 seconds in length, along with demographic information such as age groups, gender, height, weight, and pregnancy status. Out of the 1,568 patients, 60% (equivalent to 942 individuals) were provided for training. A patient could have heart sound recordings from up to six different recording locations, with a total of 5,272 recordings in the dataset. The possible locations of the heartsound recordings were the pulmonary valve, aortic valve, mitral valve, tricuspid valve, or an unspecified location. Furthermore, each patient is assigned two tags, one indicating the presence, absence, or uncertainty of heart murmurs, and the other one indicating whether the recordings contain normal or abnormal heart sound recordings. About 13% of the given data contained missing values in the metadata, which most commonly occurred concurrently in the age, height, and weight features (cf. Figure 1).

The recordings were methodically sampled by [4] using various algorithms to detect and define the primary heart sounds and their respective boundaries. Labels were assigned for data sections that cardiac physiologists deemed to be representative of high-quality segments. The remaining signal may comprise both low and high-quality data. In their research, [4] sampled signals at 4KHz, since oversampling notably beyond the Nyquist limit (double the highest frequency of the intended signal) does not offer extra insights about the signal [4]. Moreover, the heart sound signals are normalised within the [-1, 1] range.

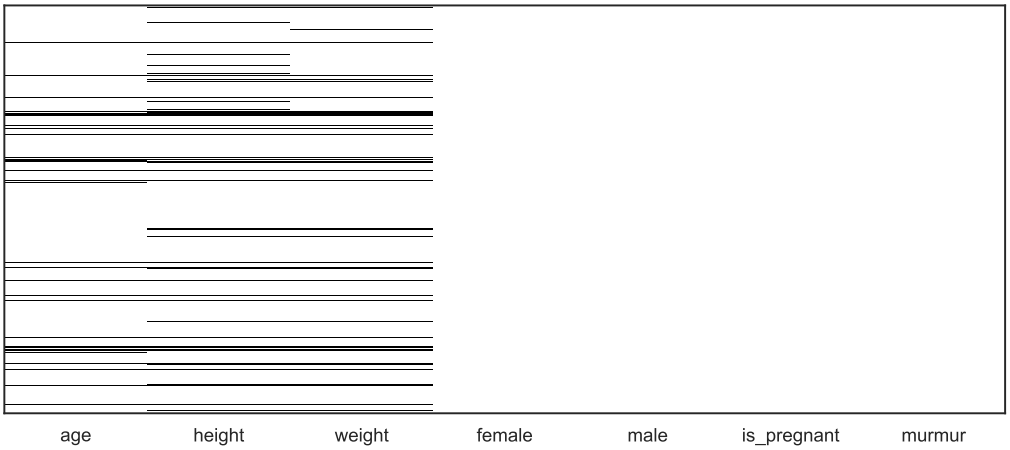

Fig 1. Missing values per demographic feature per patient (black: missing).

3.2 Multi-site data and model evaluation

In our original study we used a stratified split into training and test data (80/20). We used the murmur and outcome labels to balance both sets. The idea was to model closely the population distribution in both sets. For a more comprehensive evaluation of model robustness, in here, we employed ten-fold cross-validation. As stated in the introduction, this study does not concentrate on a specific loss function in order to maintain the generalisability of our results. Instead, the focus lies on general metrics such as the Area Under the Curve (AUC), accuracy, error rates, and confusion matrix.

Evaluating models in a multi-site context is imperative for ensuring their safe and effective implementation in real-world settings. As discussed in Section 2.2, the PhysioNet 2016 Challenge database [29] will be used for an out-of-distribution evaluation.

The database's heart sound recordings were procured from numerous contributors worldwide, collected in both clinical and non-clinical environments from healthy

individuals and patients with heart diseases. The Challenge's training set includes a total of 3,153 recordings, each lasting between 5 and 120 seconds. The records are from different body locations, typically the aortic, pulmonic, tricuspid, and mitral areas. They are categorised as either normal (79%) or abnormal (21%), with abnormal ones coming from patients with confirmed cardiac diagnoses, including heart valve defects and coronary artery disease. These recordings involve both children and adults, with each subject contributing between one and six heart sound recordings. All recordings were re-sampled to 2,000Hz and are provided in .wav format. Regrettably, the database does not allow for the linking of multiple locations to a single patient. Subject identifiers are available only for 490 recordings out of 3,153. We hence decided to treat each recording as an individual patient with one recording. For zero-shot performance we used the whole dataset for evaluation, for the fine-tuned evaluation we applied a 70/30 split.

3.3 Data preparation and feature extraction

We built and extended on the data preparation from [7]: A short-time, windowed Fourier transformation is used to derive the frequency and phase component of segments of a signal as it changes over time [64]. These features are represented as a spectrogram, which is an image depicting the change in amplitude (or power) of various frequency components over time. Owing to its effectiveness in a range of recent audio classification tasks, the spectrograms use a logarithmic mel scale for the frequency, which aims to preserve the perceived distance between pitches by humans (cf. Figure 2) [6, 65]. We performed this extraction using the SCIPY and LIBROSA Python libraries. The recordings were segmented into overlapping sections using a window of 4 seconds and a stride of 1 second. The spectrogram of each section was computed using a Fast Fourier Transform with a periodic Hanning window of 25 milliseconds, a stride of 10 milliseconds, a minimum frequency of 10Hz, and a maximum frequency of 2000Hz.

The demographic information were processed following the guidelines set by the organisers of the Challenge [6]. This processing step included converting age categories to their approximate equivalent in months, applying one-hot encoding to gender data, and transforming pregnancy status into a binary format. We addressed missing data by using mean imputation. The features extracted from the signals encompassed summary characteristics in the time and frequency domains, along with summary measures for spectral centroid, rolloff, and bandwidth.

3.4 Models

3.4.1 Pipeline architecture

Figure 3 presents a stylised representation of the data and model pipeline: classification of the individual spectrograms per location, aggregation of these classifications across locations, and a multimodal integration of the demographic data and signal features via XGBoost [66]. The individual spectrograms' classifications were aggregated using the arithmetic mean.

3.4.2 (Bayesian) Neural Network

For the classification of spectrograms, we explored two versions of our deep learning model. The first was a standard ResNet50 [67], which has been shown to be very effective in audio-related tasks [68]. This model acted as our baseline. The second model was an approximate Bayesian Neural Network (BNN) with the same architecture as the baseline model. We refer to the second model as a Binary Bayesian ResNet

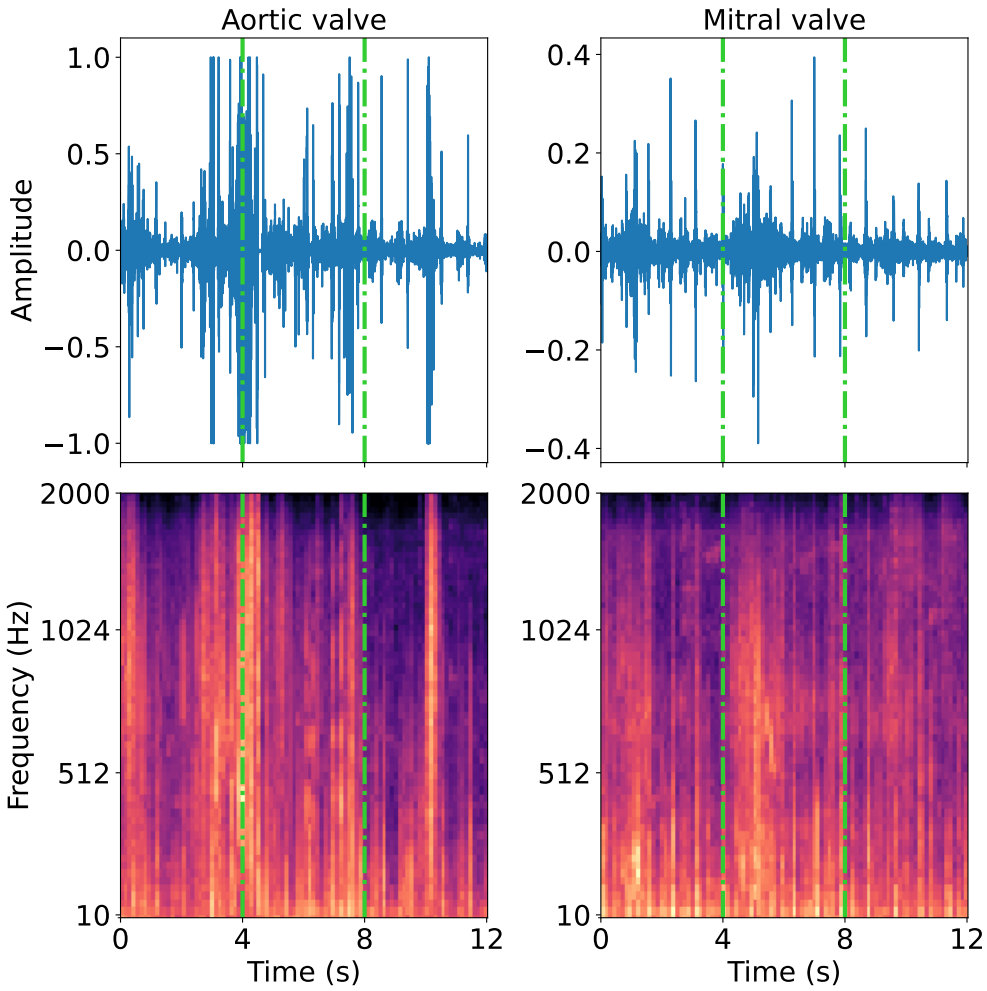

Fig 2. "Example heart sound recordings (top row) for a patient with *present* murmur recorded at the aortic valve (left column) and mitral valve (right column). The bottom row shows the log mel spectrogram, as parameterised in the code. The dash-dotted lines show how the data was partitioned into 4 second two-dimensional inputs." From [7].

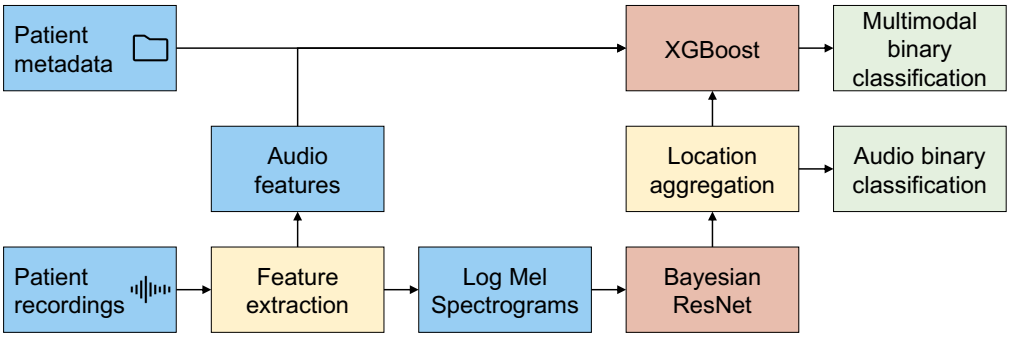

Fig 3. A schematic diagram of our data and model pipeline. Colour-coding: Blue = Data, Yellow = Fixed methods, Red = Trainable models, Green = Output.

(BBR). Both models were initialised using weights obtained from pre-training the ResNet50 on the ImageNet dataset [69].

The parameters of a BNN are distributions, instead of fixed values. This means the same input can produce diverse outputs, due to the randomness in the model parameters. BNNs built on the work of stochastic neural networks, which use either stochastic activations or weights to essentially create an ensemble of models, thereby providing a distribution over outputs and a measure of uncertainty [70]. It has been shown that BNNs can reduce overfitting, which is especially beneficial for small datasets like the one examined in this Challenge [70]. However, since this approach does not consistently outperform the deterministic counterparts, as discussed by [71], our aim was to scrutinise the specific impact of this approach in our study.

Constructing a complete BNN requires modelling the prior distribution over all model parameters, a task that can be computationally demanding [72]. On the other side, dropout layers are a common component of modern neural networks, which during each forward pass choose a random subset of the neurons to be disabled. Typically, the dropout layers are removed during inference. However, it has been shown that including them during training and inference is a viable approximation to a complete BNN [70, 72]. Inspired by [72], we added dropout layers to various segments of the ResNet50 architecture, particularly to the `BasicBlock()` and `Bottleneck()` modules, as per the ResNet implementation from [71]. This can be interpreted as a Monte Carlo approximation to BNNs. In this context, the term 'Monte Carlo' signifies the use of random sampling to generate numerical outcomes, specifically creating diverse neural network configurations via the selective deactivation of neurons. (It is worth noting that this approach does not strictly approximate BNNs but should still assist in combating overfitting. Further details on the specifics of this approximation are discussed in [72].)

3.5 NASSS evaluation framework

To provide an indication of the prospects for scaling up automated heart murmur detection, we employed the NASSS framework for an indicative assessment. The NASSS framework [10] has been developed to investigate the challenges associated with the implementation of technologies in healthcare, focusing on the risks of **N**onadoption, **A**bandonment, **S**cale-up, **S**pread, and **S**ustainability. It offers a qualitative guide comprising 19 questions, each of which can be categorised as either simple, complicated, or complex. These questions span seven dimensions: the condition or illness, the technology itself, the value proposition, the system of adopters, the organisational setting, the broader context, and the process of embedding and adaptation over time.

In addition to presenting our findings, we contextualise them by comparing with other digital healthcare technologies. Various studies have applied the NASSS framework [73] across different contexts. However, in LMIC, such as the study of wearable health monitors in Cambodia [74], the framework has been predominantly used for qualitative evaluation, omitting the quantitative assessment (simple, complicated, complex). To ensure clarity and avoid ambiguities in our analysis, we selected studies that include this quantitative dimension. Short of matching studies to ours, the examples include the assessment of telehealth consultations in Australia [75], the adoption of digital twins in healthcare [76], and the implementation of in-hospital malnutrition screening systems [77], all drawn from varied healthcare settings.

4 Results: Case study

4.1 Preliminary data analysis of training data

The provided 2022 Challenge dataset [4] is predominantly comprised of pediatric cases and reveals a noteworthy imbalance for the murmur labels. As per Table 2 and 3, 74% of the patients manifested no heart murmurs, compared to 19% who did. In a minor portion (7%) of the instances, the murmur status remained ambiguous. The outcome label is rather balanced, with 52% of the provided 942 patients being labelled as normal and 48% being labelled as abnormal. As shown in Figure 4, the distributions of age, weight, and height generally conform to expected patterns, with a few outliers (cf. Figure 4).

Table 2 shows a correlation between the occurrence of heart murmurs and abnormal clinical outcomes. However, not all instances of abnormal outcomes can be attributed to heart murmurs, suggesting that other factors also contribute.

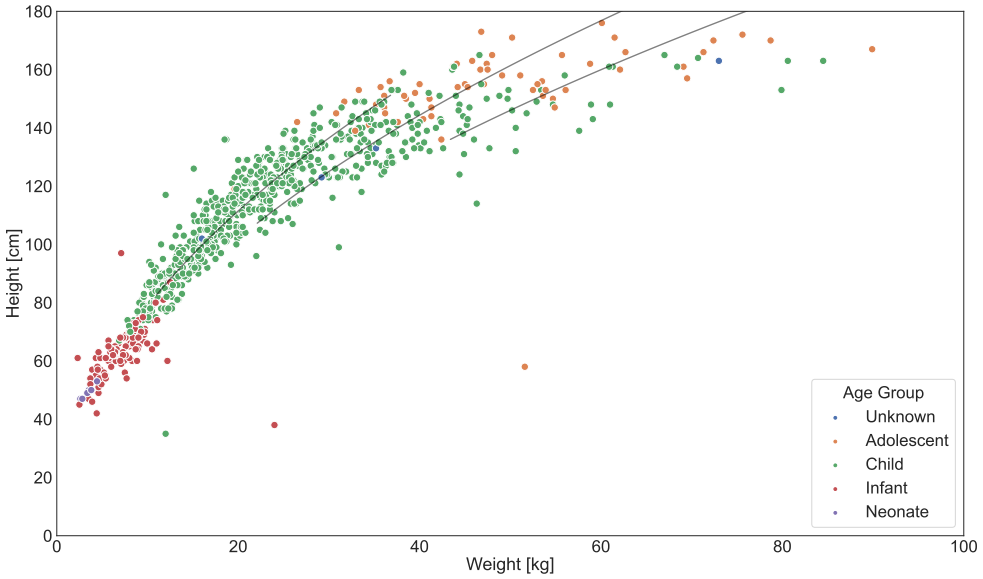

Fig 4. Distribution of age, weight, and height among patients in the training data (n=942). Age categories are as follows: Neonate, from birth to 27 days; Infant, from 28 days to 1 year; Child, from 1 to 11 years; Adolescent, from 12 to 18 years; Young Adult, from 19 to 21 years. Black lines indicate height-to-weight combinations corresponding to the medians of the median body mass indices (BMI) as proxy for a healthy BMI, within the 10th to 90th percentile weight range for the three age groups [2,8), [8,14), [14,20]. Data are derived from US sources as cited in [78], owing to its availability. $Height_{[m]} = \sqrt{Weight_{[kg]}/BMI}$.

Table 2. Murmur labels by outcome labels [n (% of column)].

	Absent	Unknown	Present	Sum
Normal	432 (62.2)	25 (36.8)	29 (16.2)	486 (51.6)
Abnormal	263 (37.8)	43 (63.2)	150 (83.4)	456 (48.4)
Sum	695	68	179	942

Table 3. Murmur labels by age [n (% of 942)].

	Absent	Unknown	Present	Sum
Neonate	4 (0.4)	1 (0.1)	1 (0.1)	6 (0.6)
Infant	76 (8.1)	25 (2.7)	25 (2.7)	126 (13.4)
Child	495 (52.6)	37 (3.9)	132 (14.0)	664 (70.5)
Adolescent	53 (5.6)	3 (0.3)	16 (1.7)	72 (7.6)
Missing	67 (7.1)	2 (0.2)	5 (0.5)	74 (7.9)
Sum	695 (73.8)	68 (7.2)	179 (19.0)	942 (100)

4.2 Preliminary data analysis of multi-site data

The multi-site evaluation employed the PhysioNet 2016 Challenge data [29, 79]. The data is unbalanced; out of the total 3,153 records, only 665 (21%) are classified as abnormal, with the rest (79%) being classified as normal. Gender information is available for 2,689 individuals, 8% of whom are female. Age data are present for 2,199 individuals, and range from 10 to 90 years, with an average age of 30. However, only 31 records include both, height and weight data. Additional demographic information, such as Body Mass Index (BMI), smoking status, and disease severity, are available for distinct subgroups of patients. Data on the location of the recording, the patient's condition, and diagnosed diseases are also occasionally available. As shown in Figure 5, 'Abnormal' recordings are on average (25.6 sec) significantly ($p < 0.001$) longer than 'Normal' recordings (21.7 sec).

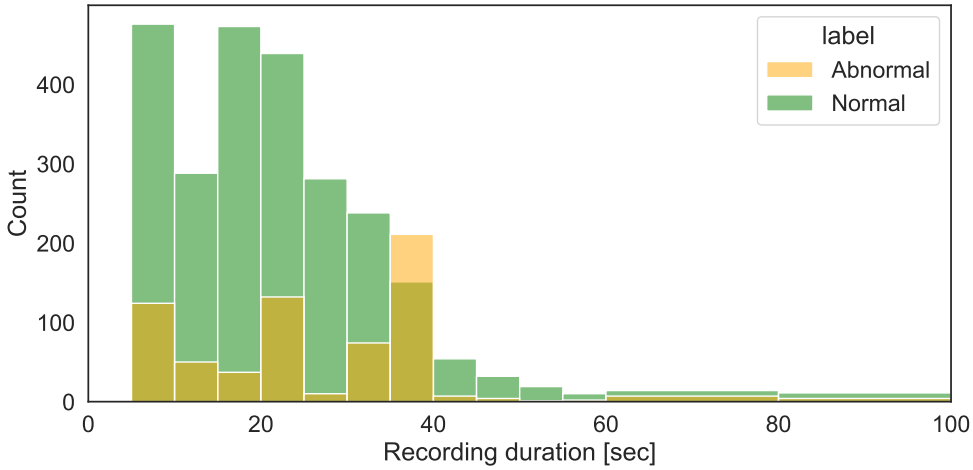

Fig 5. Distribution of recording lengths by findings in the 2016 Challenge data (Normal: n = 2488, average recording length = 21.7 sec; Abnormal: n = 665, average recording length = 25.6 sec).

4.3 Model performances

Table 4 presents an overview of the performance metrics for the various models examined in this paper.

For comparison, our initial model, designed for three-class murmur classification and termed as DBRes, achieved a weighted accuracy of 0.771 (placing it in 4th position) on the hidden test set provided by PhysioNet, and a slightly higher accuracy of 0.780 when evaluated on a locally held-out, stratified subset of the data. The congruence between

the murmur Challenge scores, obtained from our reserved portion of the training set, and those of the hidden test set from PhysioNet, suggests that our approach is well-constructed to enable the model to effectively generalise across diverse datasets.

When we compared our Bayesian approximation model (BBRes) with a pure ResNet model (Res), we observed a clear improvement across all reported metrics (see Table 4). To isolate the effect of adding dropout layers during training from that of retaining them during inference in the Bayesian approach, we also compared the results with a ResNet model where dropout layers were active only during training (Res with dropout). The results indicate that the Bayesian approach still outperforms the approach with only dropouts. (Due to the large standard deviation across splits, none of the differences proved to be statistically significant at a threshold of $p<0.01$.)

The binary models used in here consistently demonstrate accuracies and AUC values above 80% for the murmur classification. However, the performance of the outcome model is markedly subpar, with an overall accuracy that fails to surpass 60%. For context, the top ten teams in the Challenge achieved an average accuracy of 0.5662 on the hidden test data for the outcome task, with a standard deviation of 0.0159.

As illustrated in Table 5 and Table 6, both prediction tasks exhibit substantial error rates, with a pronounced inclination toward false-negative predictions for the presence of abnormalities.

Table 4. Average performance and standard deviation on ten-fold cross-validation subsets of the training set for various models. In bold best performing model. BBRes means Binary Bayesian ResNet as described under Section 3.4. DBRes was our original, multi-class model. Res is the counterpart of BBRes without the Bayesian adjustments.

Murmur models	Acc. Present/Unknown	Acc. Absent	Overall Accuracy	AUC
DBRes binary of multiclass prediction	**0.7030 (0.1821)**	0.7947 (0.0717)	0.7770 (0.0743)	0.8180 (0.1236)
Res	0.4395 (0.1739)	0.9466 (0.0714)	0.8151 (0.0749)	0.8195 (0.1194)
Res with dropout during training	0.4556 (0.2662)	0.9233 (0.0626)	0.8197 (0.0332)	0.8303 (0.0699)
BBRes	0.5033 (0.1823)	**0.9563 (0.0488)**	0.8408 (0.0614)	0.8430 (0.1381)
BBRes with XGBoost	0.5185 (0.1914)	0.9490 (0.0438)	0.8398 (0.0592)	0.8379 (0.0540)
BBRes with XGBoost, weighted	0.6268 (0.0838)	0.9526 (0.0199)	**0.8594 (0.0239)**	**0.8436 (0.0426)**
Outcome models	Acc. Abnormal	Acc. Normal	Overall Accuracy	AUC
BBRes Outcome	0.4403 (0.1042)	0.7525 (0.0893)	0.5976 (0.0653)	0.6536 (0.1041)

Table 5. Confusion matrix of our best recordings only model (BBRes), at a decision threshold of 0.5, evaluated on one randomly selected 10% held-out set. AUC=0.915, FNR=0.32.

	True present + True unknown	True absent
Pred. Present/Unknown	17	0
Pred. Absent	8	70

Table 6. Confusion matrix of our best performing, unbalanced model for the outcome label task, recordings only (BBRes Outcome), at a decision threshold of 0.5, evaluated on one randomly selected 10% held-out set. AUC=0.728, FNR=0.468.

	True Abnormal	True Normal
Pred. Abnormal	25	10
Pred. Normal	22	38

4.4 Model robustness

Implementing a model in a practical setting requires establishing a decision threshold and formulating rules based on this threshold. As demonstrated in Figure 6, the models exhibit a high sensitivity to decision thresholds. Determining an optimal action point, such as issuing a warning for a follow-up screening, presents a complex challenge. Striking a judicious balance among various types of errors is essential for ensuring the model's reliability and effectiveness in real-world applications. (An alternative approach based on ranking all predictions, rather than using a threshold, could entail directing patients with the highest scores to further screenings. However, this strategy would neither be fair to patients nor represent an efficient allocation of resources.)

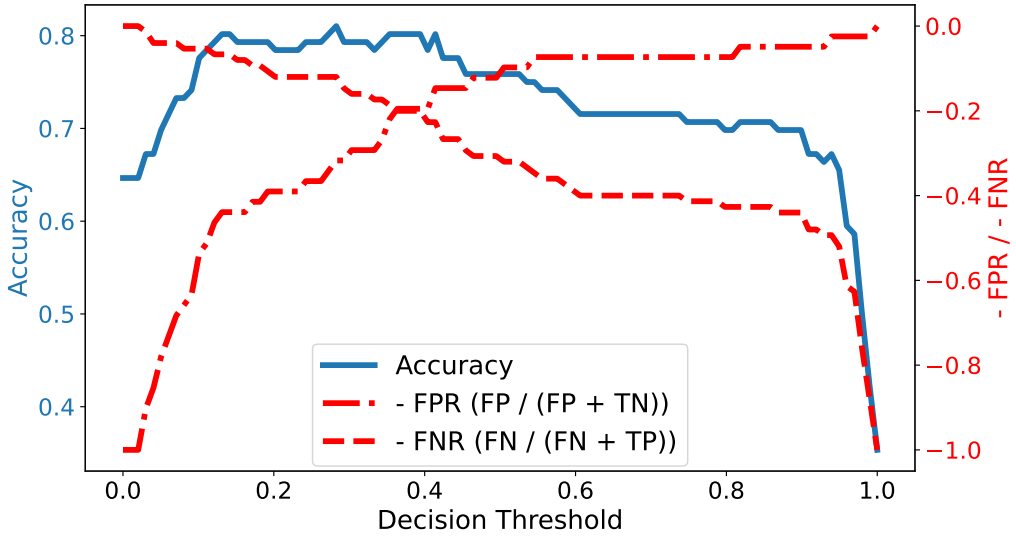

Fig 6. Accuracy, false-positive-rate and false-negative-rate of 'Present' label for different decision thresholds of the best performing binary model.

To investigate the stability of our model on out-of-distribution cases, we evaluated our model's performance on the outcome task by leveraging multi-site data from the 2016 PhysioNet Challenge in a zero-shot fashion. (For context, the 2016 Challenge's leading team, [80], reported a validation accuracy of 95.2%.) Our findings, displayed in Table 7, show a significant decline in performance when the model is applied to out-of-distribution data. The findings clearly indicate that the deployment of pre-trained models in isolation is unfeasible. To achieve robust performance, one must either standardise the data collection procedure or develop more resilient models, potentially through strategies such as improved feature extraction. (However, our results warrant cautious interpretation for several reasons: firstly, our model did not exhibit strong performance on in-distribution data (cf. Table 4); secondly, the data were sourced from adult populations rather than from children, as in the 2022 data; and thirdly, the limitation exists of having only one recording per patient.)

For further comparison, we also assessed the model from [81], notable for its reported accuracy of 99.7% and availability of published code. Originally designed for the Yaseen dataset's [30] clean, single-heartbeat recordings, this model was tested on the 2016 Challenge data without prior adjustments in a zero-shot fashion. The original approach by [81] takes the first 2 seconds of a given recording, which we adjusted to applying a majority voting over all 2 second segments of a recording. The zero-shot application resulted in an AUC of 0.501, with an accuracy notably lower than the accuracy achieved by simply labelling all outcomes as normal (cf. Table 8).

Table 7. Confusion matrix of our best performing, unbalanced model for the outcome label task, recordings only, at a decision threshold of 0.5, evaluated on the 2016's Challenge data (n=3,153) in a zero-shot fashion. AUC=0.511, Accuracy=0.521, FNR=0.472.

	True Abnormal	True Normal
Pred. Abnormal	351	1197
Pred. Normal	314	1291

Furthermore, the model's sensitivity (TPR) for detecting abnormalities was only 10% at a decision threshold of 0.5. To be clear, we do not want to criticise the study. In fact, we only picked it because it was one of the few studies that actually published their code to make their results reproducible. We simply want to highlight the difficulties and gaps of transferring models to real world applications.

Table 8. Confusion matrix of the model by [81] at a decision threshold of 0.5, evaluated on the 2016's Challenge data (n=3,153) in a zero-shot fashion. AUC=0.501, Accuracy=0.073, FNR=0.100.

	True Abnormal	True Normal
Pred. Abnormal	20 (141)	64 (45)
Pred. Normal	180 (59)	682 (701)

4.5 Deployment challenges (NASSS)

By integrating the results of our case study with those from our literature overview, we applied the NASSS framework [10] to evaluate the key challenges in deploying AI-supported heart murmur detection in low-income settings. Although a complete evaluation of dimensions 5-7 (organisation, context, and adaptation over time) is not feasible without specific knowledge of the target organisation, Figure 7 indicates that the challenges in the first four dimensions are not highly complex. Regarding the first dimension of NASSS (1A-B), the condition itself (cf. Section 3.5), heart murmurs represent a well-understood medical condition, as described in Section 1, albeit with variations in occurrence, diagnosis, and treatment across different income settings. Concerning the technology (2A-D) and its adoption (4A-C), [4] demonstrated that the system can function as a straightforward plug-and-play model requiring minimal staff training. The primary risk identified in this study pertains to the technology's dependability across different sites, operators, and systems. Ensuring standardised data collection through training and quality checks is critical. Furthermore, while the technology is desirable for patients, its financial viability hinges on the specific healthcare organisation within the target country (3A-B).

For comparison, we also present results from other studies (cf. Figure 8). Although these results must be interpreted with caution due to differences in settings, they suggest that, although a study may be assessed as relatively straightforward in the initial dimensions, it can encounter complex and complicated challenges related to the adopter and the organisational system.

In summary, we conclude that the implementation of AI-supported heart murmur detection is feasible under three main conditions: the predictive models must be robust, the organisational framework must facilitate a sustainable and scalable roll-out including follow-up care options for patients, and secure funding must be in place.

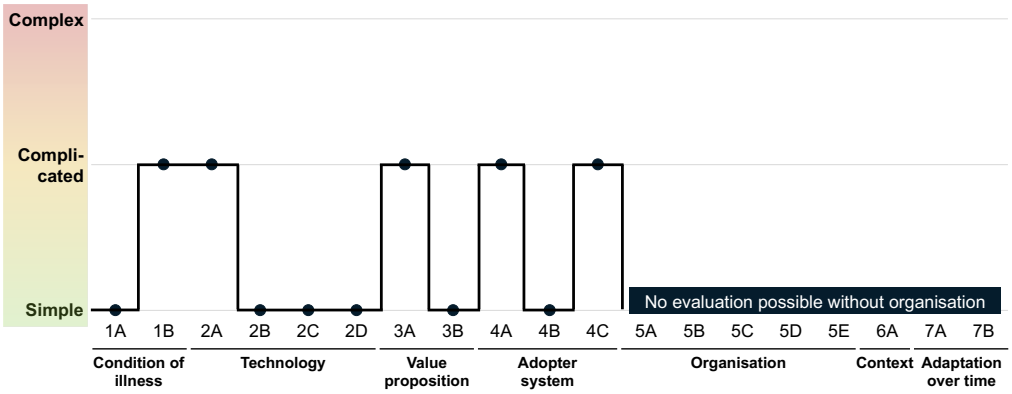

Fig 7. Indicative NASSS evaluation for the deployment of heart sound recordings in low-resource settings. (More detailed information in Table 9 in the Supporting Information Section.)

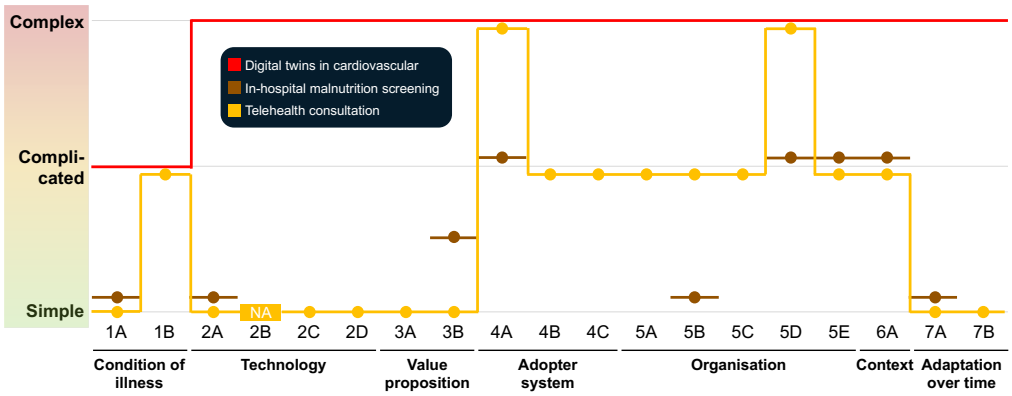

Fig 8. NASSS evaluation for various examples found in the literature from HIC settings. Red: Digital twins in cardiovascular medicine (a direct matching to the sub-categories was not possible) [76]; Brown: In-hospital malnutrition screening [77]; Orange: Telehealth consultation [75].

5 Discussion

5.1 Murmur task performance

In this research, we present and evaluate a deep learning technique for identifying murmurs and general irregularities through the analysis of heart sound recordings and demographic information. Our main approach, referred to as BBRes (originally named DBRes, or Dual-Bayesian-ResNet, in previous work [7]), employs a Binary Bayesian ResNet50 architecture to classify murmurs based on segmented spectrograms of heart audio recordings. This Bayesian model shows marked improvement over a standard ResNet architecture. The extended approach integrates results from BBRes with additional attributes derived from audio signals and patient demographics, using XGBoost for classification. Table 4 demonstrates that spectrograms are an effective data representation and, in combination with ResNet, contribute significantly to predictive performance. The inclusion of demographic information and signal features further improves overall accuracy.

The results of this study demonstrate the effectiveness of using deep neural networks for the categorisation of heart murmurs through the analysis of cardiac sound data. By enhancing the specificity of these models in identifying murmurs, they could play a key role in the development of computational screening methods for congenital heart disease.

5.2 Outcome task performance

As highlighted in Section 4.3, the performance of our model, as well as that of most Challenge models, exhibits low accuracy in the outcome task. This discrepancy is particularly striking when contrasted with results presented in recent literature (cf. Section 1.3). However, we argue that the findings from existing studies [8, 9, 81] are not directly comparable to the 2022 Challenge. For example, the Yaseen dataset [30]—often cited for models with accuracy rates exceeding 99%—features extremely short (<4 sec) and clean recordings. In contrast, the Challenge data comprise a variety of noises and longer recording durations.

To investigate these observations further, we tested our model on the Yaseen dataset and achieved an accuracy of 99%. Notably, this result was achieved without any hyperparameter tuning or model adjustments; we simply trained and tested the model on Yaseen splits. As described in Section 4.4, in a reverse experiment, we also evaluated the Yaseen model from [81] on the Challenge data. The analysis highlights the challenges in transferring AI models across different datasets in healthcare: Although the model from [81] exhibited exceptional performance on the Yaseen dataset, its efficacy significantly diminished when applied to the 2016 Challenge data. This observation is critical in understanding the limitations of AI models in healthcare, where data heterogeneity is common.

Consequently, our findings point towards the necessity for enhanced focus on data pre-processing, cleaning, robust feature extraction, and standardisation in future research. This could prove instrumental in augmenting the cross-site applicability of AI models, ensuring more robust and generalisable healthcare solutions.

5.3 Limitations and future modelling research

We identify two major limitations of our approach: A) Our models were trained exclusively on children's data, while the out-of-distribution evaluation set predominantly features adult data. B) As demonstrated throughout this paper, developing a model without considering its practical deployment proves unproductive. The choice of the correct loss function for optimisation is highly contingent on the deployment setting,

and various loss functions warrant investigation [82]. Nevertheless, we anticipate that the insights gained from this work will help in identifying gaps that require attention for successful model deployment.

Moreover, we adopted a methodology focused on directly predicting the target variable using deep learners. This strategy yielded success in the murmur task challenge by emulating the complexities of the weighted, multiclass problem. However, the approach has considerable limitations, notably in model robustness and interpretability. Alternative methodologies based on robust feature engineering, such as segmentation, have been explored by other leading teams [83]. Such approaches of robust feature engineering offer potential improvements in interpretability and may enhance model robustness against overfitting [84].

Future research could also investigate strategies to integrate patient demographic information, signal characteristics, and BBRes outputs more effectively. Additionally, the exploration of multiple fusion techniques may improve model performance [85]. We fused the features at a relatively late stage; however, an earlier feature fusion could better align with how clinicians use demographic information when interpreting charts [84]. Besides multimodal approaches, there are many other possible paths to explore to increase model robustness. Given the variability in model performance, experiments with models designed for greater robustness, such as foundation models, could be beneficial [86]. The evaluation of foundation models in healthcare applications remains an open area of research [87]. Medical foundation models like BiomedGPT [88] and Med-PaLM M [89] have yet to be tested for tasks similar to ours. Another avenue worth exploring is the replacement of the Fourier method in spectrogram creation with a signature-based approach [90]. The application of self-supervised learning to incorporate more domain-specific data shows promise, as evidenced by a recent paper that introduced HeartBEiT, a vision-based transformer model for ECG analysis [91]. HeartBEiT demonstrated significantly superior performance at lower sample sizes compared to standard CNNs. For an extensive overview of recent developments in heart sound analysis, the work by [92] offers valuable insights.

5.4 Advancing towards broad deployment

For a widespread adoption of automated pre-screening technologies, such as the one studied, several key factors require attention.

One is the implementation of a comprehensive data-mining pipeline, as shown in Figure 9. Such a pipeline typically encompasses several steps: problem comprehension, data understanding, data preparation, model training/fitting, evaluation, and deployment (cf. CRISP-DM: [93]). The process should be regarded as cyclical rather than linear to allow for continual refinement. During the initial stage of the problem comprehension, an exhaustive risk assessment proves essential for successful integration of deep learning into sensitive systems. This assessment must encompass the identification of relevant subgroups and potential data correlations. Regular evaluations and monitoring post-deployment contribute to risk minimisation and to successful employment of deep learning applications in sensitive environments.

There are a number of key aspects to prioritise, in addition to the aforementioned pipeline stages.

A) The provision of a user-friendly tool that guarantees reliable data collection and includes a quality check of the data, which is crucial for success, as indicated by [4].

B) A detailed plan for training operators to collect accurate data and pre-screen patients for eligibility [4, 48]. This plan should specify the methods, timing, and locations for screenings.

C) The assessment of the tool's adaptability to new environments without local fine-tuning, as well as the validation of models on local, representative datasets concerning population and data quality [50, 94]. This should include adaptability to variations in background noise and data collection devices [4].

D) The establishment of a well-defined communication protocol for prediction certainty [17].

E) The implementation of continuous quality monitoring to facilitate timely interventions should performance decline, along with defining appropriate metrics, which is fundamental for maintaining standards [13, 92].

F) Finally, the introduction of a clear action plan. This ensures that patients understand subsequent steps and that follow-up support is assured.

Regarding C), recent literature has started to explore automated correction for variations in data acquisition, such as when different hardware or software are used. Unsupervised alignment methods are one of the proposed solutions to address this issue [95].

By synthesising the literature [16, 17, 33, 92, 96] and the findings presented above, Figure 9 offers an overview of the considerations important for deployment. It is crucial that these steps are considered not only during deployment but also throughout the initial problem assessment and the entire process of data collection and modelling.

5.5 Ripple-effect risks

The implementation of AI solutions, particularly in sensitive domains, can induce a so-called *ripple-effect*, wherein the introduction of a new technology leads to unanticipated changes in the behavior of an existing system [97]. This necessitates a thorough understanding of how experts alter their decision-making process when interacting with AI-based systems.

In this context, two distinctive behavioural patterns have been observed, often contingent on prior experiences with such systems and their perceived reliability [98]. Firstly, *algorithm aversion* alludes to the propensity to dismiss AI systems' recommendations, often due to a lack of trust or past experiences with erroneous outcomes. Conversely, *automation bias* is the tendency to unquestioningly follow AI systems' suggestions, potentially overlooking human judgement and intuition.

These behavioral trends underscore the importance of maintaining a 'human-in-the-loop' approach in the development of AI applications for sensitive decision-making processes [98]. Addressing these tendencies can contribute to more seamless integration of AI systems into sensitive environments, thereby mitigating negative impacts and enhancing the overall effectiveness of decision-making processes.

5.6 Practical deployment challenges in low-income settings

The deployment of point-of-care (POC) monitoring technologies, such as heart murmur detection, in low-income settings is full with challenges that span from resource allocation and supply chain barriers to technical and data collection hurdles:

Limited resources pose a significant obstacle to the implementation of such technologies. This is particularly prevalent in regions where infrastructure is inadequate, ranging from electricity to internet penetration, which are both essential for the operation of POC devices [36]. The effective implementation of POC technologies relies heavily on the availability of a well-maintained supply chain, which is often lacking in low-income settings. Moreover, even with a robust supply chain, the logistical challenges

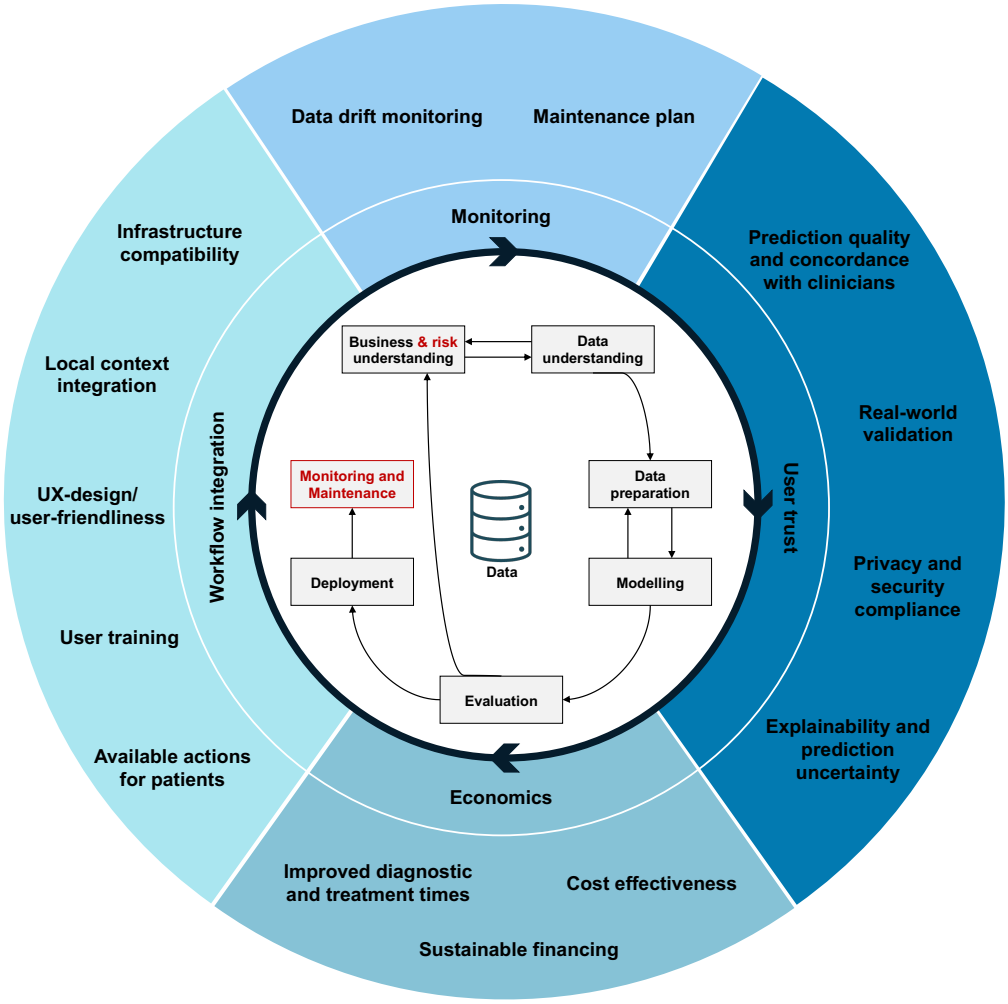

Fig 9. In blue: Important considerations for a human-centred development. Centre: Important steps for a successful deployment of AI technologies in healthcare around CRISP-DM, Cross-Industry Standard Process for Data Mining (cf. [93], in red font our additions to the original process definition).

of maintaining and updating complex technological systems in these settings can be formidable.

The cost of deploying POC technologies is another notable challenge, particularly in low-income settings. These costs can be associated with data acquisition and preparation, hardware and computing resources, as well as system maintenance and upgrading [36]. Although AI technologies have the potential to reduce costs in the long run, the initial financial investment can be a considerable barrier.

Data collection and management pose yet another significant hurdle. The process of collecting high-quality, usable data is resource-intensive, and this can be particularly challenging in low-income settings where the necessary infrastructure may be lacking. Protecting sensitive data is a paramount concern, and robust protocols must be in place to ensure that data are handled ethically and securely [46]. While AI models can be trained on globally available data, the necessity of local fine-tuning can create additional challenges. For instance, the need to gather and incorporate locally relevant data can further strain already limited resources. It also raises questions about the inclusivity and fairness of technologies, particularly in regions where data availability is limited [16, 99].

In order to ensure the successful deployment of POC technologies in low-income settings, an evidence-based approach should be employed in decision-making and implementation [36]. This includes conducting thorough risk assessments, considering the unique challenges and limitations of each setting, and prioritising sustainable, long-term solutions that can be integrated into existing systems. Ethical considerations, such as the fair and secure use of AI applications, must also be at the forefront of these efforts [16, 46, 99]. Furthermore, there are socio-cultural factors that may hinder the adoption of technologies. Therefore, solutions should focus on integrating intelligence into existing systems and institutions rather than attempting to replace them or build from scratch [36].

Besides overcoming all those challenges, a collaborative ecosystem is important for the success of AI applications in health, including a regulatory framework that provides principles and standards for data governance and a sustainable financing. Open source frameworks present an important step to lower barriers [100]. One of the biggest barriers currently is that data collection and storage are too fragmented and inaccessible [33]–a problem that HIC and LMIC share. Evidence has shown that a human-centred approach is important for the success of tools [48, 50]. A development design in which all stakeholders are considered and ethnographic fieldwork is conducted including front-line healthcare workers, such as community health workers, is important [48].

Data and code availability

Our code is available on a GitHub repository [101]. The training data are publicly available under `https://moody-challenge.physionet.org/2022/#data` and the complete collection process and data analysis of the whole dataset is described in [4]. The 2016's Challenge validation data are available under `https://physionet.org/content/challenge-2016/1.0.0/`.

Acknowledgement

Benjamin Walker was funded by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). Felix Krones was supported in part through the Friedrich Naumann Foundation by the German tax payer.

References

1. World Health Organisation. World Health Statistics; 2023.
2. World Health Organisation. Cardiovascular diseases (CVDs) — who.int; 2021. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
3. World Health Organisation. Noncommunicable diseases; 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.
4. Oliveira J, Renna F, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics. 2021;26(6):2524–2535.
5. Frank JE, Jacobe KM. Evaluation and management of heart murmurs in children. American Family Physician. 2011;84(7):793–800.
6. Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al.. Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022; 2022.
7. Walker B, Krones F, Kiskin I, Parsons G, Lyons T, Mahdi A. Dual Bayesian ResNet: A Deep Learning Approach to Heart Murmur Detection. Computing in Cardiology. 2022;.
8. Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. Deep Learning Methods for Heart Sounds Classification: A Systematic Review. Entropy. 2021;23(6). doi:10.3390/e23060667.
9. Dwivedi AK, Imtiaz SA, Rodriguez-Villegas E. Algorithms for Automatic Analysis and Classification of Heart Sounds–A Systematic Review. IEEE Access. 2019;7:8316–8345.
10. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, Hinder S, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. Journal of Medical Internet Research. 2017;19(11):e8775.
11. Kirch DG, Petelle K. Addressing the physician shortage: the peril of ignoring demography. JAMA. 2017;317(19):1947–1948.
12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44–56.
13. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. The British Journal of Radiology. 2023; p. 20220878.
14. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. The American Journal of Surgical Pathology. 2018;42(12):1636.
15. Nam JG, Hwang EJ, Kim J, Park N, Lee EH, Kim HJ, et al. AI improves nodule detection on chest radiographs in a health screening population: a randomized controlled trial. Radiology. 2023; p. 221894.

16. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature Medicine. 2022; p. 1–8.

17. Tran D, Liu J, Dusenberry MW, Phan D, Collier M, Ren J, et al. Plex: Towards reliability using pretrained large model extensions. arXiv:220707411. 2022;.

18. Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, et al. Robust and Efficient Medical Imaging with Self-Supervision. arXiv:220509723. 2022;.

19. Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, et al. EHR foundation models improve robustness in the presence of temporal distribution shift. Scientific Reports. 2023;13(1):3767.

20. Mittelstadt B, Wachter S, Russell C. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default. arXiv:230202404. 2023;.

21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453.

22. Zietlow D, Lohaus M, Balakrishnan G, Kleindessner M, Locatello F, Schölkopf B, et al. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 10410–10421.

23. Yoon JS, Oh K, Shin Y, Mazurowski MA, Suk HI. Domain Generalization for Medical Image Analysis: A Survey. arXiv:231008598. 2023;.

24. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, et al. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Translational Vision Science & Technology. 2020;9(2):45–45.

25. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello CP, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. npj Digital Medicine. 2023;6(1):111.

26. Lu C, Chang K, Singh P, Pomerantz S, Doyle S, Kakarmath S, et al. Deploying clinical machine learning? Consider the following... arXiv:210906919. 2021;.

27. Han R, Acosta JN, Shakeri Z, Ioannidis J, Topol E, Rajpurkar P. Randomized Controlled Trials Evaluating AI in Clinical Practice: A Scoping Evaluation. medRxiv. 2023; p. 2023–09.

28. MacPherson P, Webb EL, Kamchedzera W, Joekes E, Mjoli G, Lalloo DG, et al. Computer-aided X-ray screening for tuberculosis and HIV testing among adults with cough in Malawi (the PROSPECT study): A randomised trial and cost-effectiveness analysis. PLOS Medicine. 2021;18(9):e1003752.

29. Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In: 2016 Computing in Cardiology Conference (CinC); 2016. p. 609–612.

30. Yaseen, Son GY, Kwon S. Classification of Heart Sound Signal Using Multiple Features. Applied Sciences. 2018;8(12).

31. Bentley P, Nordehn G, Coimbra M, Mannor S. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results; 2011. http://www.peterjbentley.com/heartchallenge/index.html.

32. Hoodbhoy Z, Hasan B, Siddiqui K. Does artificial intelligence have any role in healthcare in low resource settings. Journal of Medical Artificial Intelligence. 2019;2(13):10–21037.

33. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: a systematic scoping review. npj Digital Medicine. 2022;5(1):162.

34. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Global Health. 2018;3(4):e000798.

35. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MY, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. The Lancet Digital Health. 2019;1(1):e35–e44.

36. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial intelligence for healthcare in Africa. Frontiers in Digital Health. 2020;2:6.

37. Moftakhar L, Mozhgan S, Safe MS. Exponentially increasing trend of infected patients with COVID-19 in Iran: a comparison of neural network and ARIMA forecasting models. Iranian Journal of Public Health. 2020;49(Suppl 1):92.

38. Tiwari S, Kumar S, Guleria K. Outbreak trends of coronavirus disease–2019 in India: a prediction. Disaster Medicine and Public Health Preparedness. 2020;14(5):e33–e38.

39. Buscema M, Asadi-Zeydabadi M, Lodwick W, Nde Nembot A, Bronstein A, Newman F. Analysis of the ebola outbreak in 2014 and 2018 in West Africa and Congo by using artificial adaptive systems. Applied Artificial Intelligence. 2020;34(8):597–617.

40. Nakasi R, Tusubira JF, Zawedde A, Mansourian A, Mwebaze E. A web-based intelligence platform for diagnosis of malaria in thick blood smear images: A case for a developing country. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 984–985.

41. Aguiar FS, Torres RC, Pinto JV, Kritski AL, Seixas JM, Mello FC. Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. Medical & Biological Engineering & Computing. 2016;54:1751–1759.

42. Young C, Barker S, Ehrlich R, Kistnasamy B, Yassi A. Computer-aided detection for tuberculosis and silicosis in chest radiographs of gold miners of South Africa. The International Journal of Tuberculosis and Lung Disease. 2020;24(4):444–451.

43. Cao Y, Liu C, Liu B, Brunette MJ, Zhang N, Sun T, et al. Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. In: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE; 2016. p. 274–281.

44. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmology. 2019;137(10):1182–1188.

45. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmology. 2019;137(9):987–993.

46. Sallstrom L, Morris O, Mehta H. Artificial intelligence in Africa's healthcare: Ethical considerations. ORF Issue Brief. 2019;312.

47. Arun C. AI and the Global South: Designing for other worlds. The Oxford Handbook of Ethics of AI. 2019;.

48. Okolo CT. Optimizing human-centered AI for healthcare in the Global South. Patterns. 2022; p. 100421.

49. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. 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; 2020. p. 1–12.

50. Widner K, Virmani S, Krause J, Nayar J, Tiwari R, Pedersen ER, et al. Lessons learned from translating AI from development to deployment in healthcare. Nature Medicine. 2023; p. 1–3.

51. Ismail A, Kumar N. AI in global health: The view from the front lines. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–21.

52. Love SM, Berg WA, Podilchuk C, López Aldrete AL, Gaxiola Mascareño AP, Pathicherikollamparambil K, et al. Palpable breast lump triage by minimally trained operators in Mexico using computer-assisted diagnosis and low-cost ultrasound. Journal of Global Oncology. 2018;4:1–9.

53. Kisling K, Zhang L, Simonds H, Fakie N, Yang J, McCarroll R, et al. Fully automatic treatment planning for external-beam radiation therapy of locally advanced cervical cancer: a tool for low-resource clinics. Journal of Global Oncology. 2019;5:1–9.

54. Garzon-Chavez D, Romero-Alvarez D, Bonifaz M, Gaviria J, Mero D, Gunsha N, et al. Adapting for the COVID-19 pandemic in Ecuador, a characterization of hospital strategies and patients. PLOS ONE. 2021;16(5):e0251295.

55. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. New England Journal of Medicine. 2023;388(21):1981–1990.

56. AI W. AI for Social Impact - Wadhwani AI — wadhwaniai.org; 2023. `https://www.wadhwaniai.org/`.

57. Aidoc. Aidoc Always On Healthcare AI — aidoc.com; 2023. `https://www.aidoc.com/`.

58. AI U. Ubenwa - giving hope to newborns — ubenwa.ai; 2023. `https://www.ubenwa.ai/`.

59. OpenMRS. OpenMRS.org — openmrs.org; 2023. `https://openmrs.org/`.

60. DHIS2. OpenMRS.org — openmrs.org; 2023. `https://dhis2.org/`.

61. Mollura DJ, Culp MP, Pollack E, Battino G, Scheel JR, Mango VL, et al. Artificial intelligence in low-and middle-income countries: innovating global health radiology. Radiology. 2020;297(3):513–520.

62. Lind Plesner L, Müller FC, Brejnebøl MW, Laustrup LC, Rasmussen F, Nielsen OW, et al. Commercially available chest radiograph AI tools for detecting airspace disease, pneumothorax, and pleural effusion. Radiology. 2023;308(3):e231236.

63. MidMeds. 3M Littmann 3200 Electronic Stethoscope: Black; 2023 [cited 2023-11-26]. Available from: `https://www.midmeds.co.uk/littmann-electronic-3200-stethoscope-black-p-4263.html`.

64. Sejdić E, Djurović E, Jiang J. Time–frequency feature representation using energy concentration: An overview of recent advances. Digital Signal Processing. 2009;19:153–183.

65. Wisdom S, Erdogan H, et al. DCASE 2021 Task 4: Sound event detection and separation in domestic environments; 2021.

66. Pimentel MAF, Mahdi A, Redfern O, Santos MD, Tarassenko L. Uncertainty-aware model for reliable prediction of sepsis in the ICU. In: 2019 Computing in Cardiology (CinC); 2019. p. 1–4.

67. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2016. p. 770–778.

68. Palanisamy K, Singhania D, Yao A. Rethinking CNN models for audio classification. arXiv:200711154. 2020;.

69. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 248–255.

70. Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M. Hands-on Bayesian neural networks—A tutorial for deep learning users. IEEE Computational Intelligence Magazine. 2022;17:29–48.

71. Kiskin I, Sinka M, Cobb AD, Rafique W, Wang L, Zilli D, et al. HumBugDB: a large-scale acoustic mosquito dataset. arXiv:211007607. 2021;.

72. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR; 2016. p. 1050–1059.

73. Shin HD, Hamovitch E, Gatov E, MacKinnon M, Samawi L, Boateng R, et al. The NASSS (Non-Adoption, Abandonment, Scale-Up, Spread and Sustainability) framework use over time: A scoping review. medRxiv. 2023; p. 2023–11.

74. Liverani M, Ir P, Perel P, Khan M, Balabanova D, Wiseman V. Assessing the potential of wearable health monitors for health system strengthening in low- and middle-income countries: a prospective study of technology adoption in Cambodia. Health Policy and Planning. 2022;37(8):943–951. doi:10.1093/heapol/czac019.

75. Cartledge S, Rawstorn JC, Tran M, Ryan P, Howden EJ, Jackson A. Telehealth is here to stay but not without challenges: a consultation of cardiac rehabilitation clinicians during COVID-19 in Victoria, Australia. European Journal of Cardiovascular Nursing. 2022;21(6):548–558.

76. Winter PD, Chico TJ. Using the Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) Framework to Identify Barriers and Facilitators for the Implementation of Digital Twins in Cardiovascular Medicine. Sensors. 2023;23(14):6333.

77. Besculides M, Mazumdar M, Phlegar S, Freeman R, Wilson S, Joshi H, et al. Implementing a Machine Learning Screening Tool for Malnutrition: Insights From Qualitative Research Applicable to Other Machine Learning–Based Clinical Decision Support Systems. JMIR Formative Research. 2023;7(1):e42262.

78. Fryar CD, Carroll MD, Gu Q, Afful J, Ogden CL. Anthropometric reference data for children and adults: United States, 2015-2018. National Center for Health Statistics Vital Health Stat 3(46). 2021;.

79. Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement. 2016;37(12):2181.

80. Maknickas V, Maknickas A. Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. Physiological Measurement. 2017;38(8):1671.

81. Nguyen MT, Lin WW, Huang JH. Heart Sound Classification Using Deep Learning Techniques Based on Log-mel Spectrogram. Circuits, Systems, and Signal Processing. 2023;42(1):344–360.

82. Cobb AD, Roberts SJ, Gal Y. Loss-calibrated approximate inference in Bayesian neural networks. arXiv:180503901. 2018;.

83. McDonald A, Gales MJ, Agarwal A. Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models. In: 2022 Computing in Cardiology (CinC). vol. 498. IEEE; 2022. p. 1–4.

84. Duvieusart B, Krones F, Parsons G, Tarassenko L, Papież B, Mahdi A. Multimodal Cardiomegaly Classification with Image-Derived Digital Biomarkers. In: Medical Image Understanding and Analysis; 2022. p. 13–27.

85. Krones F, Walker B, Parsons G, Lyons T, Mahdi A. Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest. Computing in Cardiology. 2023;50:Preprint.

86. Bommasani R, Hudson DA, et al. On the opportunities and risks of foundation models. arXiv:210807258. 2021;.

87. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering. 2022; p. 1–7.

88. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks. arXiv:230517100. 2023;.

89. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards Generalist Biomedical AI. arXiv:230714334. 2023;.

90. Morrill J, Fermanian A, Kidger P, Lyons T. A generalised signature method for multivariate time series feature extraction. arXiv:200600873. 2020;.

91. Vaid A, Jiang J, Sawant A, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. npj Digital Medicine. 2023;.

92. Ren Z, Chang Y, Nguyen TT, Tan Y, Qian K, Schuller BW. A Comprehensive Survey on Heart Sound Analysis in the Deep Learning Era. arXiv:230109362. 2023;.

93. IBM. CRISP-DM; 2021. Available from: `https://www.ibm.com/docs/it/spss-modeler/saas?topic=dm-crisp-help-overview`.

94. Mitchell WG, Dee EC, Celi LA. Generalisability through local validation: overcoming barriers due to data disparity in healthcare. BMC Ophthalmology. 2021;21(1):1–3.

95. Roschewitz M, Khara G, Yearsley J, Sharma N, James JJ, Ambrózay É, et al. Automatic correction of performance drift under acquisition shift in medical image classification. Nature Communications. 2023;14(1):6608.

96. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Annals of Translational Medicine. 2020;8(7).

97. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019. p. 59–68.

98. De-Arteaga M, Fogliato R, Chouldechova A. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020. p. 1–12.

99. Partnership A. Artificial Intelligence for Africa: An Opportunity for Growth, Development, and Democratisation; 2014.

100. Abhinav V, Krisstina R, Vivek E, Yukti S. Building a collaborative ecosystem for AI in healthcare in Low and Middle Income Economies. Atlantic Council GeoTech Center. 2020;.

101. Walker B, Krones F, Kiskin I, Parsons G, Lyons T, Mahdi A. Dual Bayesian ResNet: A Python code for heart murmur detection. GitHub repository; 2022. `https://github.com/Benjamin-Walker/PhysionetChallenge2022`.

Supporting Information

Table 9. Indicative NASSS evaluation for the deployment of heart sound recordings in low-resource settings (cf. [10]).

Domain	Question		Rating	
The condition or illness	What is the nature of the condition or illness?	1A	1	Well-characterized, well-understood, predictable
The condition or illness	What are the relevant sociocultural factors and comorbidities?	1B	2	Must be factored into care plan and service model
The technology	What are the key features of the technology?	2A	2	Not yet developed or fully interoperable; not 100% dependable
The technology	What kind of knowledge does the technology bring into play?	2B	1	Directly and transparently measures [changes in] the condition
The technology	What knowledge and/or support is required to use the technology?	2C	1	None or a simple set of instructions
The technology	What is the technology supply model?	2D	1	Generic, "plug and play" solutions requiring minimal customization; easily substitutable if supplier withdraws
The value proposition	What is the developer's business case for the technology (supply-side value)?	3A	2	Business case underdeveloped; potential risk to investors
The value proposition	What is its desirability, efficacy, safety, and cost effectiveness (demand-side value)?	3B	1	Technology is desirable for patients, effective, safe, and cost effective
The adopter system	What changes in staff roles, practices, and identities are implied?	4A	2	Existing staff must learn new skills and/or new staff be appointed
The adopter system	What is expected of the patient and is this achievable by, and acceptable to them?	4B	1	Nothing
The adopter system	What is assumed about the extended network of lay caregivers?	4C	2	Assumes a caregiver will be available when needed
The organization	What is the organization's capacity to innovate?	5A	na	na
The organization	How ready is the organization for this technology-supported change?	5B	na	na
The organization	How easy will the adoption and funding decision be?	5C	na	na
The organization	What changes will be needed in team interactions and routines?	5D	na	na
The organization	What work is involved in implementation and who will do it?	5E	na	na
The wider context	What is the political, economic, regulatory, professional (eg, medicolegal), and sociocultural context for program rollout?	6A	na	na
Embedding and adaptation over time	How much scope is there for adapting and coevolving the technology and the service over time?	7A	na	na
Embedding and adaptation over time	How resilient is the organization to handling critical events and adapting to unforeseen eventualities?	7B	na	na