Towards Efficient Annotations for a Human-AI Collaborative, Clinical Decision Support System: A Case Study on Physical Stroke Rehabilitation Assessment

Min Hun Lee
mhlee@smu.edu.sg
Singapore Management University
Singapore

Daniel P. Siewiorek
dps@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Asim Smailagic
asim@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal

Sergi Bermúdez i Badia
sergi.bermudez@m-iti.org
University of Madeira, NOVA-LINCS
Funchal, Portugal

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) algorithms are increasingly being explored to support various decision-making tasks in health (e.g. rehabilitation assessment). However, the development of such AI/ML-based decision support systems is challenging due to the expensive process to collect an annotated dataset. In this paper, we describe the development process of a human-AI collaborative, clinical decision support system that augments an ML model with a rule-based (RB) model from domain experts. We conducted its empirical evaluation in the context of assessing physical stroke rehabilitation with the dataset of three exercises from 15 post-stroke survivors and therapists. Our results bring new insights on the efficient development and annotations of a decision support system: when an annotated dataset is not available initially, the RB model can be used to assess post-stroke survivor's quality of motion and identify samples with low confidence scores to support efficient annotations for training an ML model. Specifically, our system requires only 22 - 33% of annotations from therapists to train an ML model that achieves equally good performance with an ML model with all annotations from a therapist. Our work discusses the values of a human-AI collaborative approach for effectively collecting an annotated dataset and supporting a complex decision-making task.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **User studies**; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03.
https://doi.org/10.1145/3490099.3511112

KEYWORDS

Human Centered AI; Human-AI Collaboration; Human-In-the-Loop Systems; Clinical Decision Support Systems; Physical Stroke Rehabilitation Assessment;

ACM Reference Format:
Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2022. Towards Efficient Annotations for a Human-AI Collaborative, Clinical Decision Support System: A Case Study on Physical Stroke Rehabilitation Assessment. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3490099.3511112

1 INTRODUCTION

Recent advances in artificial intelligence (AI) have made it applicable to support decision-making in healthcare [6, 12, 14, 42]. Specifically, researchers have explored the feasibility of clinical decision support systems that analyze a large amount of data using AI techniques and provide supplementary information as a secondary set of perspectives to enhance the accuracy and efficiency of a clinician's decision making [13].

Clinical decision support systems can be categorized as either a rule-based system [3, 10, 41, 55] or a machine learning-based system [6, 12, 22, 39, 53], depending on its data analytic techniques. A rule-based system requires the engagements with clinicians (e.g. semi-structured interviews) to capture and translate their knowledge into a set of rules [41, 55]. This rule-based system has the benefit of being interpretable and flexible. However, it remains challenging to elicit a global set of rules that represent the knowledge of clinicians on a complex decision-making task [41, 49]. In contrast to a rule-based system that relies on inputs from clinicians, a machine learning (ML)-based system utilizes an ML algorithm to automatically extract patterns from annotated data on a decision-making task. Specifically, significant recent research has been devoted to exploring a deep neural network in clinical applications (e.g. cancer diagnosis [20] and rehabilitation assessment [39, 48]). As researchers have demonstrated that an ML-based system is competent to perform a task that requires clinician expertise [20, 39], researchers have also explored the feasibility to integrate this system into practice.

Although ML-based systems have shown competitive performance, prior work describes the difficulty to integrate these systems into practice due to the lack of human-centered designs as a primary reason for failure [13, 31, 40, 60] and performing as a black-box system [52]. To address this latter issue, researchers have actively explored a method or tool to provide explanations on the predictions of an ML-based system [25, 40, 43, 50]. In addition, there is growing human-computer interaction research that involves the end-users to understand their practices and needs [13, 40, 60] and socio-environmental factors [6] for the design and evaluation of a system. Yet, most prior work assumes to have a large amount of annotated data [24] to develop an ML-based system. There is a limited exploration on how to support an expensive process to collect annotated data.

In this paper, we present a human-AI collaborative, clinical decision support system that combines a machine learning (ML) model with a rule-based (RB) model as a hybrid model (HM) to assist the assessment of a post-stroke patient's quality of motion (Figure 1). We contribute to empirical research that brings insights on the development process of a system and the value of RB and HM models to support efficient annotations for training an ML model.

As an ML model typically requires a large amount of annotated, training data [24], this system first leverages an RB model from therapists to assess the quality of patient's rehabilitation exercises and identify samples with low confidence scores. Instead of annotating all samples, a therapist can review quantitative assessment on a patient's exercise from the system (Figure 2) and annotate samples with low confidence scores. Once the annotated dataset is collected, this system trains an ML model and integrates it with a rule-based model as a hybrid model. This hybrid model can not only derive generic insights from data with an ML model, but also provide an opportunity for user engagement with an RB model for human-AI collaborative decision-making [42].

For the evaluation, we utilized the dataset of three upper-limb rehabilitation exercises from 15 post-stroke patients and 11 healthy participants [39]. Given this dataset, we compared the training of a machine learning (ML) model (i.e. a neural network) through leave-one-patient-out cross-validation using the full annotations from a therapist and the annotations from our human-AI collaborative, clinical decision support (HAC-CDS) system. For the annotations of our HAC-CDS system, we developed a rule-based (RB) model with 15 independent *if-then* rules from the interviews with therapists [41]. We then recruited five therapists to annotate patient's exercises with low confidence scores from the RB model. Our experimental results show that our system requires only 22 - 33% of total annotations (i.e. an average of 103 - 140 annotations) to train an ML model that achieves similar performance with an ML model with the full annotations.

Overall, this work advances ongoing discussions around human-AI collaborative decision making in high-stakes domains (e.g. health) [13, 42, 43]. Specifically, this work proposes a human-AI collaborative approach for efficient annotations to develop a clinical decision support system and contributes to its empirical evaluation in the context of physical stroke rehabilitation assessment. Our results provide new insights on the efficient development of a human-AI collaborative, decision support systems in high-stake domains: at the beginning of developing these systems, we recommend researchers engaging with domain experts to develop an interpretable, rule-based (RB) model [41, 43, 49] and utilize it to support efficient annotations for training an ML model. In addition, after collecting an annotated dataset, we discuss the potential of integrating an ML model with an RB model as a hybrid model to support human-AI collaborative decision making in high-stake domains [42].

2 RELATED WORK

2.1 Opportunities & Challenges of Decision Support Systems

Machine learning (ML) algorithms are increasingly being explored to support human decision-making in high-stakes situations, such as public services, criminal justice, and health [6, 12, 14, 18, 30, 36]. Following prior research that describes the potential of ML algorithms to outperform human experts on prediction tasks [17, 20, 34, 39], researchers have investigated the feasibility of deploying these ML-based decision support systems to improve decision makings. However, the development and integration of these systems in practice remain a challenge due to the lack of human-centered designs and performing as a "black-box" system [11–13, 17, 21, 31, 40, 60].

To support understanding of the predictions of an ML-based system, researchers have actively explored a method [8, 33, 40, 50] or a visualization tool [25] to provide explanations on the prediction of an ML-based system. In addition, there is increasing recent research efforts [6, 13, 21, 40, 54, 60] that highlight the importance of involving stakeholders to understand their practices and needs [13, 40, 60] and socio-environmental factors [6] for the design and evaluation of a system. Yang et al. conducted a field evaluation on the design of a clinical decision support tool for cardiologists with synthetic data and found that clinicians are more likely to embrace a tool that augments their decision-making in natural and intuitive ways [60]. Lee et al. [40] conducted interviews and focus-group sessions with therapists to understand the challenges and needs during rehabilitation assessment to design a human-centered, clinical decision support system. In addition, researchers discuss the necessity of evaluating a system in socio-technical contexts [6, 21], and creating ongoing feedback loops with stakeholders [54]. As the predictions of machine learning algorithms cannot be perfect [12], De-Arteaga et al. highlighted the importance of making a human-in-the-loop pipeline to avoid harmful effects from erroneous algorithmic recommendations instead of relying on a fully automated approach [17].

While these prior studies provide new perspectives on the necessity of engaging the stakeholders for human-centered algorithmic decision-making [56], these studies assume to have a large amount of annotated data [24] to develop an ML-based system. There has been limited exploration of how such a human-in-the-loop pipeline can be developed and how to support an expensive process to collect annotated data. In this work, we focus on a decision-making task on physical stroke rehabilitation assessment [57, 58] and explore the development of a human-AI collaborative, clinical decision support system and its effectiveness to collect an annotated dataset.

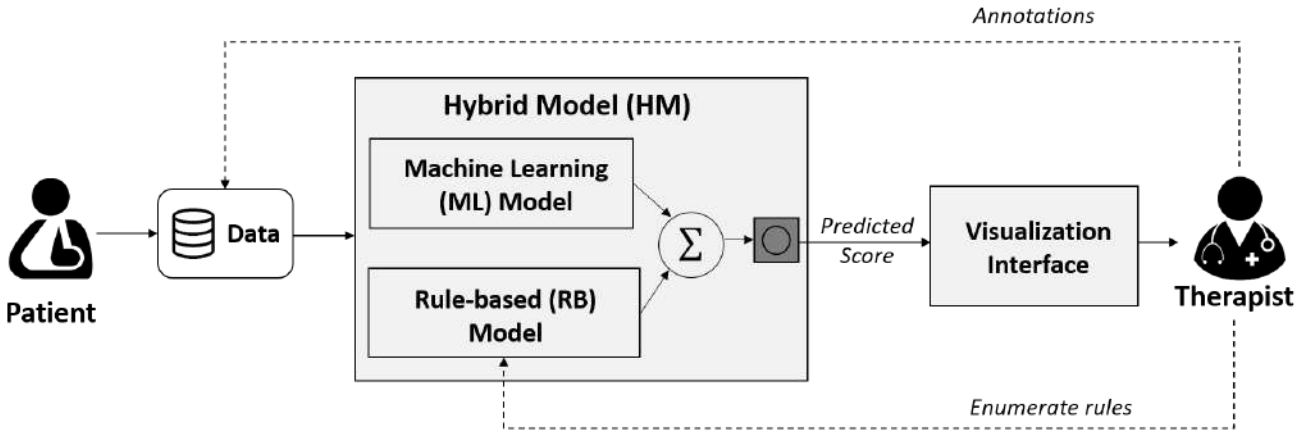


Figure 1: Flow diagram of a human-AI collaborative, clinical decision support system for physical stroke rehabilitation assessment. This system integrates a machine learning (ML) model with a rule-based (RB) model as a hybrid model to assess the quality of patient's rehabilitation exercises. After reviewing quantitative analysis on patient's exercises from the system, a therapist can provide annotations to further improve the system. Initially, this system will operate with only an RB model from therapists to facilitate the annotations. Once the annotated dataset is collected, this system will train a machine learning model and integrate it with a rule-based model as a hybrid model for human-AI collaborative decision making.

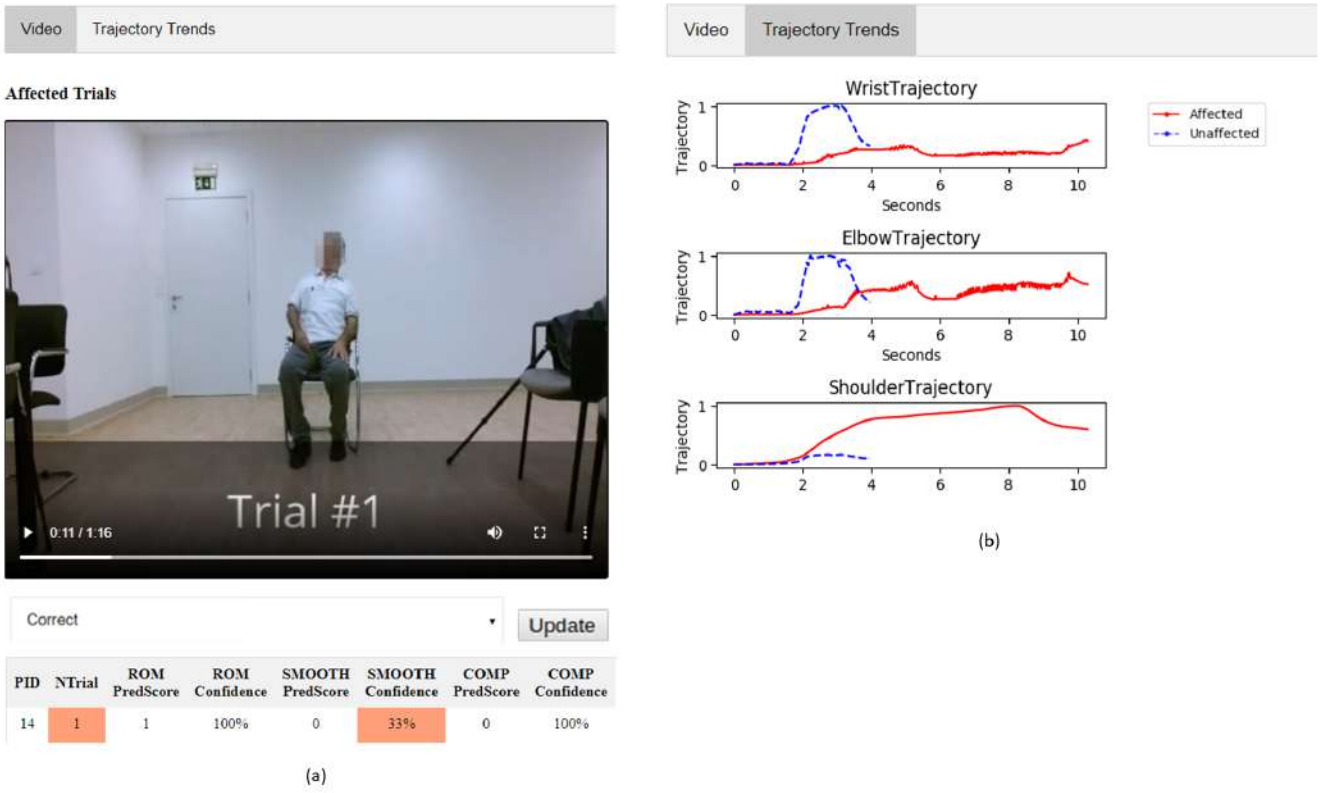


Figure 2: The visualization interface of the proposed system that presents (a) the video of patient's exercise motions and the quantitative assessment on patient's quality of motion and (b) trajectory trends between unaffected and affected sides.

2.2 Technology-Assisted Rehabilitation Assessment

Physical rehabilitation is one of the effective treatment approaches for musculoskeletal and neurological disorders (e.g. stroke) [47]. During rehabilitation, therapists utilize clinical tests that require their direct observation on patient's exercises to understand the patient's status and prescribe an intervention [57, 58]. Although rehabilitation assessment is important to determine a customized intervention of a patient, it is infrequently performed due to the therapist's limited availability [45]. Therapists often encounter difficulty with making an informed decision on patient's rehabilitation [5, 27].

Researchers have explored an approach of automatically monitoring and assessing patient's exercises with machine learning and sensors to improve current practices of rehabilitation [59]. One approach is a rule-based model, in which a set of monitoring rules is elicited from domain experts. For instance, Lee et al. compared the positions of wrist and spine joints to monitor the completion of an upper-limb rehabilitation exercise [41]. This rule-based approach can be flexible to develop a customized model. However, it remains challenging to enumerate a global set of rules to represent experts' complex decision-making processes. An alternative approach is to utilize machine learning algorithms with labeled sensor data and learn a model to assess the quality of motion [39, 48]. For example, Das et al. applied Support Vector Machine (SVM) to distinguish mild and severe symptoms of Parkinson's Disease using full-body motion capture data from four Parkinson's patients [16]. However, the development of these ML model requires an annotated dataset [24], which is labor-intensive [15].

Instead of relying on either a rule-based (RB) model or a machine learning (ML) model alone, we present a human-AI collaborative, clinical decision support system that combines an ML model with an RB model to assess the quality of post-stroke patient's rehabilitation exercises. Initially, when an annotated dataset is not available, this system operates with an RB model to predict the assessment on the quality of motion and identify samples with low confidence scores. Therapists then leveraged these confidence scores to prioritize which samples to annotate. When the annotated dataset is collected, our system trains an ML model and integrates it with the RB model to exploit the strength of both ML and RB models and support human-AI collaborative decision-making [42].

Human-in-the-loop, interactive ML approaches have been actively explored by researchers to create a better ML model with improved performance and user acceptance [1, 12, 37, 42] and improve the annotations [28, 32, 35]. Prior work shows the feasibility of presenting relevant information of a task and acquiring inputs of a user (e.g. constraints of a model [29] or feature relevance [37, 42]) to refine an ML model. In addition, researchers have discussed the importance of iterating on data (e.g. collecting new data or labels) [26] and explored interactive approaches for annotations in various applications: sound event detection [32], semantic annotations [35], and animal behavior [28]. This work contributes to the empirical research on the development of a human-AI collaborative, decision support system that integrates an ML model with an RB model and evaluation of its efficiency for annotations in the context of assessing physical stroke rehabilitation exercises.

3 STUDY ON PHYSICAL STROKE REHABILITATION ASSESSMENT

This work focuses on the application of assessing physical stroke rehabilitation exercises, which is a major part of patient care for stroke, a common and disabling global healthcare problem [38]. In this section, we describe the designs of our study to assess physical stroke rehabilitation that includes three upper-limb exercises and kinematic features to represent the quality of motion.

3.1 Three Task-Oriented Upper Limb Exercises

This work utilizes three upper-limb physical stroke rehabilitation exercises, recommended by therapists [39]. The first exercise (Exercise 1) is called *"Bring a cup to the mouth"*. For Exercise 1, a post-stroke survivor has to raise the post-stroke survivor's wrist to the mouth as if drinking water. The second exercise (Exercise 2) is referred to as *"Switch a Light on"*. For Exercise 2, a post-stroke survivor has to raise the post-stroke survivor's hand forward as if touching a light switch on the wall. The third exercise (Exercise 3) is *"Move a cane forward"*. For Exercise 3, a post-stroke survivor has to practice the usage of a cane by extending the post-stroke survivor's elbow in the seated position. These exercises are selected due to their correspondence with major motion patterns: elbow flexion for Exercise 1, shoulder flexion for Exercise 2, elbow extension for Exercise 3.

3.2 Performance Components and Kinematic Features

This work leverages a Kinect v2 sensor (Microsoft, Redmond, USA) to track the joint positions of a post-stroke survivor while performing an exercise and extracts various kinematic features to represent the quality of motion [39, 51, 57, 58]. Specifically, this work assesses the post-stroke survivor's quality of motion in terms of the following three performance components: the *'Range of Motion (ROM)'*, *"Smoothness"*, and *"Compensation"*.

The *"ROM"* performance component indicates whether a post-stroke survivor achieves to perform the target position of an exercise (e.g. bringing the wrist to the mouth for Exercise 1). For the *'ROM'* performance component, we computed joint angles (e.g. elbow flexion, shoulder flexion, elbow extension), normalized relative trajectory (i.e. Euclidean distances between two joints - head and wrist, head and elbow), and normalized trajectory distance (i.e. absolute distances between two joints - head and wrist, shoulder and wrist) in x, y, z-axis.

The *"Smoothness"* performance component describes whether a post-stroke survivor smoothly coordinates a motion. For the *"Smoothness"* performance component, we computed various speed-related features: speed, acceleration, jerk, the zero-crossing ratio of acceleration and jerk, and Mean Arrest Period Ratio (the portion of the frames when speed exceeds 10% of the maximum speed) [51]. As this study focuses on upper-limb exercises, we computed these speed-related features on wrist and elbow joints.

The *"Compensation"* performance component check whether a post-stroke survivor utilizes any unnecessary joint motions (e.g. leaning the trunk forward, elevating the shoulder joint, etc.) to perform an exercise. For the *'Compensation'* performance component, we computed joint angles (i.e. the elevated angle of a shoulder,

the tilted angle of spine, and shoulder abduction) and normalized trajectories (i.e. the distances between joint positions of the head, spine, shoulder joints in x, y, z-axis from the initial to the current frames).

Before extracting features, we applied a moving average filter with the window size of five frames to reduce the noise of acquiring joint positions from a Kinect sensor similar to the prior research [39]. For each exercise motion, we computed a feature matrix ($\mathbf{F} \in R^{t \times d}$) with t frame and d features and statistics (i.e. max, min, range, average, and standard deviation) over all frames of the exercise to summarize a motion into a feature vector ($X \in R^{5d}$). We denote the correct, normal performance component as $Y = 1$ and incorrect, abnormal performance component as $Y = 0$.

4 HUMAN-AI COLLABORATIVE, CLINICAL DECISION SUPPORT SYSTEM

In this work, we present a human-AI collaborative, clinical decision support system (Figure 1) that integrates a machine learning (ML) model with a rule-based (RB) model to assist the assessment of post-stroke survivor's quality of motion. Initially, this system will start with an RB model from therapists as an ML model requires an annotated dataset for the development. With the RB model, this system assesses the quality of post-stroke survivor's rehabilitation exercises and identifies samples with low confidence scores that a therapist can prioritize to annotate. When the annotated dataset is collected, this system trains an ML model and integrates it with the RB model as a hybrid model. The hybrid model can leverage the strengths of both an ML model that extracts new insights from data and an RB model that allows user engagement (e.g. updating a rule-based model [42]) and being adaptable to support human-AI collaborative decision making on rehabilitation assessment.

4.1 Machine Learning Model

A machine learning model utilizes a supervised learning algorithm to assess post-stroke survivor's quality of motion on three performance components of physical stroke rehabilitation: *"ROM"*, *"Smoothness"*, and *"Compensation"*. Among various traditional supervised learning algorithms (e.g. Decision Trees, Linear Regression, Support Vector Machine, and Neural Networks), we utilized a Neural Network (NN) due to its outperformance as shown in [39]. For the implementation of a NN model, we explored various architectures (i.e. one to three layers with 32, 64, 128, 256, 512 hidden units) and an adaptive learning rate with different initial learning rates (i.e. 0.0001, 0.005, 0.001, 0.01, 0.1) and grid searched the parameters during the leave-out-patient cross-validation. We applied *'ReLu'* activation functions on hidden units of a NN model and trained it using cross-entropy loss ($-\sum_c \mathbf{1}_{(Y=c)} \log f(\mathbf{X})_c$), where $f(\mathbf{X})_c$ indicates the probability score of a class label c from a NN model and $\mathbf{1}_{(Y=c)}$ is an indicator function (0 or 1) if class label c is the correct classification. For training the parameters of a NN model, we utilized *'AdamOptimizer'* until the tolerance of optimization became 0.0001 or the maximum 200 iterations. The parameters of neural networks (i.e. hidden layers/units and learning rate) that achieved the best F1-score during the cross-validation are summarized in Table 1.

Table 1: Parameters of machine learning models (i.e. neural network models) using full annotations by a therapist

	Hidden Layers and Units / Learning Rate		
	ROM	Smooth	Comp
E1	(32, 32, 32) / 0.1	(16) / 0.0001	(256, 256) / 0.1
E2	(256) / 0.1	(512, 512) / 0.1	(128) / 0.1
E3	(256) / 0.1	(64, 64) / 0.001	(128, 128) / 0.1

4.2 Rule-Based Model

A rule-based (RB) model utilizes the set of feature-based rules from therapists to assess post-stroke survivor's quality of motion. For the initial development of an RB model, we elicited 15 independent *if-then* rules to assess the quality of physical stroke rehabilitation exercises from semi-structured interviews with two therapists [41]. During the interviews, the researcher asked therapists to think out loud their procedures to quantitatively assess post-stroke survivor's quality of motion on three performance components of physical stroke rehabilitation: *"ROM"*, *"Smoothness"*, and *"Compensation"* (Section 3.2). To facilitate therapists' think-aloud procedures, the researcher showed therapists the videos of post-stroke survivors' exercises from the previous study on automated assessment of physical stroke rehabilitation exercises [39]. When therapists enumerated a particular body or joint position that they consider for assessment, the researcher described a kinematic feature to represent it and further discussed with therapists whether their processes were correctly represented by a kinematic feature. The list of 15 independent *if-then* rules from the interviews can be found in Table 2.

For assessing *"ROM"* performance component, our rules describe whether the estimated target position of each exercise is achieved or not. For example, the assessment of the ROM component for Exercise 1 was specified as follows:

$$\hat{Y} = \begin{cases} 1 & \text{if } p^{max}(wr, c_y) >= p^{max}(spsh, c_y) \\ 0 & \text{else} \end{cases} \quad (1)$$

where $\hat{Y}$ denotes the predicted label on a performance component. $p^{max}(j, c)$ indicates the maximum joint position with a joint j (e.g. the wrist (wr) and the spine shoulder, the top of spine, ($spsh$)) and the coordinate of a joint position, c in the set $C \in \{c_x, c_y, c_z\}$. This rule monitors whether the maximum position of the post-stroke survivor's wrist joint ($p^{max}(wr, c_y)$) exceeds that of the post-stroke survivor's spine shoulder joint ($p^{max}(spsh)$) in the y-coordinate to roughly assess whether a post-stroke survivor achieves the target position of Exercise 1 that requires a post-stroke survivor to bring the post-stroke survivor's wrist to the mouth as if drinking water.

For *"Smoothness"* performance component, as therapists described that they monitor whether a post-stroke survivor can smoothly move survivor's wrist, we discussed creating rules that measure the rate at which wrist accelerations on x, y, z-axis change from positive to negative or from negative to positive and represent the degree of non-smoothness.

Table 2: List of independent rules to assess the quality of motion from therapists

Performance Components	Rules
Range of Motion (ROM)	a wrist joint should be located above a spine-shoulder joint near a head joint for exercise 1 a wrist joint should be located higher than a shoulder joint for exercise 2 a wrist joint should be located further than hip near a knee for exercise 3
Smoothness	a wrist joint should be smoothly coordinated in the x-axis during 80% of the motion (zero-crossing ratio of a wrist acceleration in the x-axis is within 20%) a wrist joint should be smoothly coordinated in the y-axis during 80% of the motion (zero-crossing ratio of a wrist acceleration in the y-axis is within 20%) a wrist joint should be smoothly coordinated in the z-axis during 80% of the motion (zero-crossing ratio of a wrist acceleration in the z-axis is within 20%)
Compensation	a head joint should not be located more/less than 15% of an initial head position in the x-axis a head joint should not be located above/below 15% of an initial head position in the y-axis a head joint should not be located more/less than 15% of an initial head position in the z-axis a spine joint should not be located more/less than 15% of an initial spine position in the x-axis a spine joint should not be located above/below 15% of an initial spine position in the y-axis a spine joint should not be located more/less than 15% of an initial spine position in the z-axis a shoulder joint should not be located more/less than 15% of an initial shoulder position in the x-axis a shoulder joint should not be located above/below 15% of an initial shoulder position in the y-axis a shoulder joint should not be located more/less than 15% of an initial shoulder position in the z-axis

For *"Compensation"* performance component, therapists mainly check whether a post-stroke survivor performs any unnecessary movements on the head, spine, and shoulder joints. We discussed creating rules that quantitatively evaluate such compensatory movements by measuring how much post-stroke survivor's head, spine, and shoulder joints are moved from the initial position in x, y, z-axis.

The equation of an RB model to compute a score of being correct on the performance component is described as follows:

$$P_{RB} = \frac{1}{|\mathbb{R}|} \sum_{r \in \mathbb{R}} \min(\frac{f_r}{\tau_r}, 1) \tag{2}$$

where $\mathbb{R}$ indicates a set of rules from therapists. f_r describes the feature value of a rule r from an exercise motion (e.g. $p^{max}(wr, c_y)$ for the example above) and τ_r describes the threshold value of a rule r (e.g. $p^{max}(spsh, c_y)$ for the example above). This equation normalizes the feature value of a rule with the threshold of a rule to compute the score of being correct. min function is applied so that this equation assigns the value of 1 if the feature value of a rule exceeds the threshold of that rule.

4.3 Hybrid Model

A hybrid model integrates two perspectives on assessment using a weighted average ensemble technique [4, 41]: a machine learning model that discovers how to assess the quality of motion from data and a rule-based model from therapists. For the assessment of the quality of motion, a hybrid model (HM) computes a weighted average of prediction scores from two models, in which the contribution of each model is weighted by the performance of a model (i.e. the F1-score of each model in the range of $[0, 1]$). The equation to compute the prediction score of an HM (P_{HM}) model is described as follows:

$$P_{HM} = \frac{\rho_{ml}}{\rho_{ml} + \rho_{rb}} P_{ML} + \frac{\rho_{rb}}{\rho_{ml} + \rho_{rb}} P_{RB} \tag{3}$$

where P_{ML} and P_{RB} indicate the predicted scores of a machine learning (ML) model and a rule-based (RB) model, and ρ_{ml} and ρ_{rb} describe the performance, F1-score of an ML model and an RB model respectively.

4.4 Visualization Interface

We implemented a web-based visualization interface using HTML, Javascript, and Python libraries to present the video of a post-stroke survivor's rehabilitation exercise, the predicted assessment on the post-stroke survivor's quality of motion, and the trajectories of the post-stroke survivor's wrist, elbow, and shoulder joints (Figure 2). When this interface presents the predicted assessment, it also includes the confidence score to *"make clear how well the system can do"* [2]. Specifically, we follow prior research that leverages the probability score of a classifier as a confidence score (e.g. probabilities from the softmax layer of a neural network) [23, 46, 61] and utilize the predicted score of a model, $P_M(Y|X)$ to identify low confidence samples (e.g. 0.2) [44], where $M \in \{ML, RB\}$. When the prediction of a model has a low confidence score, this interface highlights this prediction with low confidence so that a therapist can review and update the labels if necessary. For the user study with therapists, we pre-processed post-stroke survivors' exercise videos and stored predictions on post-stroke survivor's quality of motion from our model to avoid any delays in processing data.

As therapists utilize patient's unaffected side as normality for assessment [57, 58], this interface compares patient's unaffected and affected sides on major joint trajectories of an upper-limb exercise (i.e. shoulder, elbow, and wrist joints) to *'match relevant*

social norms' [2]. In addition, this interface supports to *"honor user feedback"* [37]. Specifically, a therapist can review the predicted assessment on a performance component of patient's rehabilitation exercises and update the labels to train a machine learning model (Figure 1).

5 EXPERIMENTS

5.1 Dataset of Three Upper Limb Exercises

This work utilized the dataset of three upper limb exercises from 15 post-stroke survivors with various functional abilities (37 ± 21 Fugl Meyer Scores [57]) and 11 healthy participants [39]. This dataset of each exercise includes the videos of participants and the corresponding body joint trajectories from a Kinect v2 sensor (Microsoft, Redmond, USA). A post-stroke survivor performed 10 repetitions of each exercise with their affected and unaffected sides and a healthy participant performed 15 repetitions of each exercise with the participant's dominant side.

5.2 Annotations

We collected three sets of annotations from therapists without/with our human-AI collaborative, clinical decision support system and evaluation of its effectiveness on an annotation process. Three sets of annotations include 1) annotations from therapists, 2) annotations from a rule-based model and therapists, and 3) annotations from a hybrid model and a therapist.

5.2.1 *Annotations from Therapists.*

Two therapists (TP 1 and 2 in Table 3) individually annotated the dataset (Section 5.1) without reviewing analysis of our system (Figure 2). As therapist 1 (TP 1) supported the recruitment of patients and evaluated their functional abilities with the clinical assessment tool (e.g. Fugl Meyer Assessment [57]), we utilized the annotations of TP 1 as ground truth for the development and evaluation of a system. In addition, the annotations of therapist 2 (TP 2) were compared with those of TP 1 to measure the therapists' agreement level (Table 4).

5.2.2 *Annotations with a Human-AI Collaborative System.*

We collected two annotations with our human-AI collaborative, clinical decision support system to evaluate its effectiveness to support annotations: annotations from a rule-based model and therapists (ARBT) and annotations from a hybrid model and a therapist (AHMT).

For the ARBT, we implemented our system with a rule-based (RB) model that does not require any annotations for the development. Our system can generate the predicted assessment on the quality of motion (Figure 2), which can serve as an annotation. In addition, our system identifies a sample with a low confidence score as described in Section 4.4 and allows a therapist to relabel annotations from the system. For this re-labeling process, we recruited five therapists with $\mu = 4.00$, $\sigma = 1.67$ years of experience in stroke rehabilitation (TPs with checkmarks in the *'Relabel'* column in Table 3). Each therapist was instructed to review the predicted assessment from the system and update its predictions to make them as accurate as possible during a 30-minute session. We assigned non-overlapping, three patients for each therapist to generate annotations on our entire dataset.

For the AHMT, we trained a machine learning model (e.g. a neural network) with the ARBT and integrated it with a rule-based as a hybrid model (HM). Our system utilizes the HM to predict the assessment of the quality of motion. For the re-labeling process, our system identifies samples with low confidence scores from the HM and replaces predictions of our system with the annotations from therapist 2 (TP 2).

Table 3: List of participants for the studies on annotation, rule elicitation (ElicitRule), and re-label (Relabel).

ID	Studies			# of Years in
	Annotation	ElicitRule	Relabel	Stroke Rehab
TP1	✓	✓		6
TP2	✓	✓	✓	4
TP3				9
TP4			✓	4
TP5			✓	1
TP6			✓	6
TP7			✓	5

6 RESULTS

6.1 System Performance of Rehabilitation Assessment

For the evaluation of system implementation, we applied the Leave-One-Patient-Out (LOPO) cross-validation on post-stroke patients to avoid overfitting and provide more reliable error estimates. The LOPO cross-validation utilizes data from all participants except one post-stroke patient to train a machine learning (ML) model and utilizes data of the left-out testing post-stroke patient to test how well an ML model can assess post-stroke patient's quality of motion. For the performance metric, we utilized an F1-score, which seeks to balance between precision (i.e. how many instances a model can classify correctly) and recall (i.e. how robust a model is).

Table 4 summarizes the performance of our system, which measures the agreement with therapist 1 (TP 1)'s annotations using average F1-scores of models on three exercises. Specifically, we present the performance of our rule-based (RB) model from therapists, a machine learning, Neural Network model (ML-NN), and a hybrid model (HM) that integrates the ML-NN model with the RB model.

The RB model achieves the lowest agreement level with therapist 1 (TP 1)'s annotations: 0.5821 average F1-scores over three exercises. The machine learning, Neural Network model (ML-NN) achieves the highest, good agreement with TP 1's annotations: 0.8279 average F1-scores over three exercises. In addition, the ML-NN also performs better than the therapists' agreement. However, the performance difference between the ML-NN and therapists' agreement is not statistically significant according to the paired t-test over their performances on three exercises and three performance components. The hybrid model (HM) integrates the ML-NN with the RB model and achieves a 0.7931 average F1-score over three exercises. The performance of the HM is a 0.03 lower average F1-score than that of the ML-NN. However, the performance difference between

Table 4: Performance of machine learning models with neural networks (ML-NN), the rule-based (RB) models, and hybrid models (HMs), and therapists' agreement

	Exercise 1 (E1)	Exercise 2 (E2)	Exercise 3 (E3)	Overall
RB	0.6148 ± 0.1702	0.6932 ± 0.1630	0.4384 ± 0.1569	0.5821 ± 0.1066
ML - NN	0.8632 ± 0.0816	0.8388 ± 0.0518	0.7818 ± 0.0097	0.8279 ± 0.0605
HM	0.8437 ± 0.0697	0.7545 ± 0.0561	0.7812 ± 0.0479	0.7931 ± 0.0644
Therapists' Agreement	0.7455 ± 0.2054	0.8147 ± 0.1522	0.7254 ± 0.1838	0.7619 ± 0.1626

the HM and the ML-NN is not statistically significant using the paired t-test. Even if combining two perspectives on assessment (i.e. the ML-NN and RB models) does not improve the performance of a model to assess the quality of motion, the HM still achieves higher performance than the therapists' agreement without statistical significance.

6.2 Effect of Annotating with a Human-AI Collaborative System

To evaluate the effectiveness of annotating with our human-AI collaborative system, we utilized three annotations to train a machine learning, Neural Network model (ML-NN). Given ML-NN models with three annotations, we compared their performance to replicate therapist's assessment and their training curves over the number of annotations from therapists. Three annotations include 1) annotations from the TP 1, 2) annotations from the rule-based model and therapists (ARBT), and 3) annotations from the hybrid model and a therapist (AHMT). For the ARBT, therapists provided an average of 141 relabels out of 465 samples. In addition, our system applied an average of 103 relabels out of 465 samples with the annotations from therapist 2 (TP2) for the AHMT.

Table 5 summarizes the performances of the ML-NNs that are trained with the leave-one-patient-out cross-validation and three different annotations respectively. For comparing the performance of the ML-NNs with three different annotations, we conducted paired t-tests over their performances on three exercises and three performance components. In addition, we show the learning curves of the ML-NNs with three different annotations (Figure 3). The orange dotted lines of Figure 3 indicate the therapists' agreement using the annotations from therapist 1 and therapist 2. The green graphs of Figure 3 indicate the performances of ML-NNs that were trained with annotations from therapist 1. The red graphs of Figure 3 describe the performance of the ML-NNs with the annotations from the rule-based model and relabeling from therapists. The sky blue graphs of Figure 3 show the performance of the ML-NNs with the annotations from the hybrid model and relabeling with the annotations from the therapist 2.

Our results in Table 5 show that the ML-NNs with three annotations perform equally well according to the paired t-tests and achieve better performance than the therapists' agreement without any statistical significance. In addition, our results in Figure 3 describe that our system with the RB model requires only around 33% annotations (i.e. an average of 141 relabels) to train the ML-NNs that have equally good performance with the ML-NNs that are trained with the annotations from the TP 1. In addition, our system with the hybrid model requires only around 22% annotations (i.e. an average of 103 relabels) to train the ML-NNs that have equally good performance with the ML-NNs that are trained with the annotations from the TP 1.

7 DISCUSSION

In this section, we discussed the potential implications and limitations of our study for the better development and integration of a human-AI collaborative, clinical decision support system in practice.

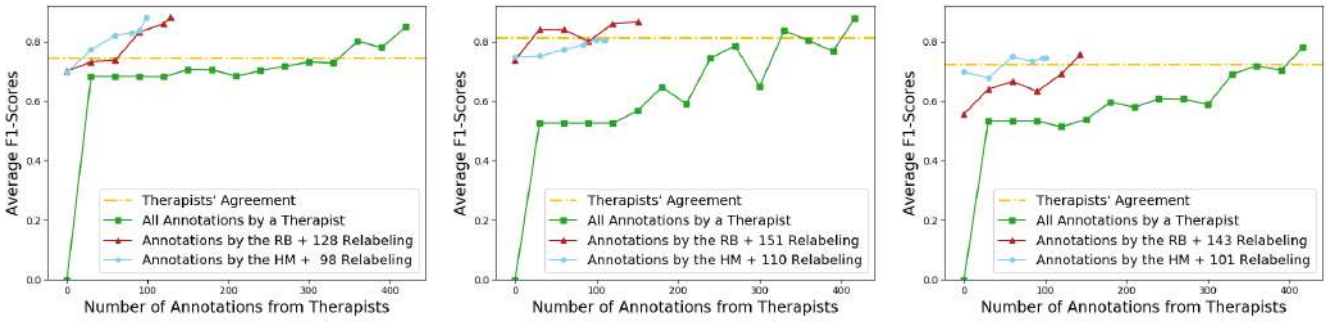
Our results suggest a possible, efficient procedure to develop a human-AI collaborative, decision support system. When an annotated dataset is not available, our system starts with a rule-based (RB) model from experts to predict expert decision-making (e.g. rehabilitation assessment) and identify samples with low confidence scores for efficient annotations (Figure 3). When an annotated dataset becomes available, our system trains an ML model to automatically extract new insights on experts' decision-making from data. Even if an ML model with a complex algorithm (e.g. a neural network) outperforms an RB model, we do not recommend replacing an RB model with an ML model that operates as a black-box system. Instead of further exploring complex, black-box algorithms to seek performance improvement [39, 48], this work suggests developing a hybrid model (HM) that integrates an ML model with an RB model for human-AI collaborative decision-making [42].

A hybrid model (HM) does not necessarily improve the performance to predict an expert's decision-making (e.g. assessing post-stroke survivor's rehabilitation exercises) as the HM integrates an ML model with an RB model that includes generic rules and performs lower than therapists' agreement. We observed that the performance of the HM became slightly lower than that of the ML model with a Neural Network (ML-NN). Such performance degradation becomes more eminent as the HM model increases the weight on the RB model that achieved lower performance than therapists' agreement (i.e. in case of exercise 1 and exercise 2). However, the HM still achieves comparable performance with the ML-NN and achieves better performance than the therapists' agreement (Table 4). Compared to the ML-NN, the HM can leverage both data analytic capability of an ML model and the flexibility and interpretability of an RB model. Specifically, the HM has the potential benefit of supporting the user engagement by analyzing rules of the RB model and fine-tuning a model with patient-specific rules for improving its performance [42].

Our results show that both RB and HM models can be utilized to identify samples with low confidence scores and support an efficient

Table 5: Performance of machine learning (ML) models with different annotation datasets: (1) annotations from a therapist, (2) annotations from the RB model and therapists, and (3) annotations from the HM model and a therapist.

	Annotations	Exercise 1 (E1)	Exercise 2 (E2)	Exercise 3 (E3)	Overall
ML - NN	a therapist	0.8632 ± 0.0816	0.8388 ± 0.0518	0.7818 ± 0.0097	0.8279 ± 0.0605
ML - NN	RB & therapists	0.8830 ± 0.0970	0.8678 ± 0.0972	0.7566 ± 0.0820	0.8358 ± 0.0690
ML - NN	HM & a therapist	0.8814 ± 0.0722	0.8052 ± 0.1222	0.7473 ± 0.0630	0.8113 ± 0.0673
Therapists' Agreement	therapists	0.7455 ± 0.2054	0.8147 ± 0.1522	0.7254 ± 0.1838	0.7619 ± 0.1626



(a) Exercise 1 **(b) Exercise 2** **(c) Exercise 3**

Figure 3: The training curve of a machine learning (ML) model (i.e. a neural network) with full annotations from a therapist (green graph), annotations from our system with only a rule-based model (red graph), and a hybrid model (skyblue graph). Our system requires only 22 - 33% of total annotations to train an ML model that achieves similar performance with an ML model with the full annotations from a therapist.

Table 6: Parameters of machine learning models (i.e. neural network models) using annotations by a rule-based model and therapists

	Hidden Layers and Units / Learning Rate		
	ROM	Smooth	Comp
E1	(64, 64, 64) / 0.1	(512, 512) / 0.1	(32, 32) / 0.1
E2	(32) / 0.0001	(256, 256) / 0.1	(16, 16) / 0.0001
E3	(64, 64) / 0.1	(64, 64, 64) / 0.1	(128, 128) / 0.1

Table 7: Parameters of machine learning models (i.e. neural network models) using annotations by a hybrid model and a therapist

	Hidden Layers and Units / Learning Rate		
	ROM	Smooth	Comp
E1	(64, 64) / 0.1	(256) / 0.1	(512, 512) / 0.01
E2	(32) / 0.0001	(256, 256) / 0.005	(16) / 0.0001
E3	(16, 16, 16) / 0.1	(128, 128, 128) / 0.005	(16) / 0.0001

annotation process (Table 5). Specifically, our system with the RB requires an average of 141 relabels, 33% of total annotations to train an ML model, and our system with the HM requires an average of 103 relabels, 22% of total annotations to train an ML model that achieves equally good performance with an ML model with full annotations by an expert. However, we found that even if the HM had much higher performance than the RB model, the ML model with the annotations from the HM and a therapist led to slightly lower performance than the ML model with the annotations from the RB model and therapists without statistical significance. Thus, it would be interesting to further explore whether the HM can support training a better ML model and achieve more efficient annotation after improving the technique of a multimodal machine learning [4] and fine-tuning the RB model. In addition, as the probability scores of a model might not be well-calibrated to represent samples with low confidence scores [23], it is important to further investigate an approach to estimate the confidence score of a model and determine if the prediction of a model can be utilized it to better identify samples with low confidence scores and build a trustful interaction with the user [7, 23].

Overall, our work presents a human-AI collaborative, clinical decision support system that exploits not only a machine learning (ML) model to automatically extract new insights from data, but also a rule-based (RB) model to accommodate experts' knowledge. We believe that a general concept of our human-AI collaborative

approach that integrates an ML model with an RB model as a hybrid model and accommodates user knowledge and inputs can be extended to other disciplines and improve their decision-making procedures. However, this study is limited to exploring the feasibility of our approach in the context of physical stroke rehabilitation assessment. An additional study is required to expand the application of our human-AI collaborative approach on different tasks or data modalities [9, 19]. In addition, as this work focused on empirical research on the development process of a system, this work provided limited interactions to support therapists' relabeling, enable a trustful usage with the system, and enhance their clinical decision-making [12, 42]. It is necessary to further study how and what additional interactions can be provided to enable trustful usage by the domain experts and improve their decision making without introducing biases [21].

8 CONCLUSION

In this work, we contributed to an empirical study on the efficient development and annotation process of a human-AI collaborative, clinical decision support system in the context of physical stroke rehabilitation assessment. Our results showed that initially if an annotated dataset is not available, a rule-based (RB) model from therapists can be used to replicate expert's decision making (e.g. assessing patient's quality of motion) and identify samples with low confidence scores to support efficient annotations for training an ML model. Specifically, our approach can leverage annotations and confidence scores of the RB model and require only 22-33% of annotations to train an ML model that achieves equally good performance with an ML model with all annotations from a domain expert. In addition, after collecting an annotated dataset, our work discusses the potential of a hybrid model that augments an ML model with an RB model from domain experts for human-AI collaborative decision making in high-stake domains.

ACKNOWLEDGMENTS

The authors thank all participants in this study for their dedication, time, and valuable inputs. Also, we thank the anonymous reviewers for their constructive comments and suggestions on the manuscript. This work is partially supported by the National Science Foundation (CNS-1518865). Additional support was provided by the IntelligentCare project (LISBOA-01-0247-FEDER-045948), the FCT LARSyS - Plurianual funding 2020-2023 (UIDB/50009/2020), and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[3] PK Anooj. 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences* 24, 1 (2012), 27–40.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[5] Mark T Bayley, Amanda Hurdowar, Carol L Richards, Nicol Korner-Bitensky, Sharon Wood-Dauphinee, Janice J Eng, Marilyn McKay-Lyons, Edward Harrison, Robert Teasell, Margaret Harrison, et al. 2012. Barriers to implementation of stroke rehabilitation evidence: findings from a multi-site pilot project. *Disability and rehabilitation* 34, 19 (2012), 1633–1638.

[6] 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.

[7] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1, 1 (2019), 20–23.

[8] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[9] Norbert Buch, Sergio A Velastin, and James Orwell. 2011. A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on intelligent transportation systems* 12, 3 (2011), 920–939.

[10] Bruce G Buchanan and Richard O Duda. 1983. Principles of rule-based expert systems. In *Advances in computers*. Vol. 22. Elsevier, 163–216.

[11] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. 2017. Unintended consequences of machine learning in medicine. *Jama* 318, 6 (2017), 517–518.

[12] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 4.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[15] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. 2019. How to develop machine learning models for healthcare. *Nature materials* 18, 5 (2019), 410.

[16] Samarjit Das, Laura Trutoiu, Akihiko Murai, Dunbar Alcindor, Michael Oh, Fernando De la Torre, and Jessica Hodgins. 2011. Quantitative measurement of motor symptoms in Parkinson's disease: A study with full-body motion capture data. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 6789–6792.

[17] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. 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*. 1–12.

[18] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[19] TT Dhivyaprabha, P Subashini, and Marimuthu Krishnaveni. 2016. Computational intelligence based machine learning methods for rule-based reasoning in computer vision applications. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.

[20] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[21] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[22] Eren Gultepe, Jeffrey P Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. 2014. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association* 21, 2 (2014), 315–325.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[24] Jatinder ND Gupta, Guisseppi A Forgionne, and Manuel Mora. 2007. *Intelligent decision-making support systems: foundations, applications and challenges*. Springer Science & Business Media.

[25] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* 25, 8 (2018), 2674–2693.

[26] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[27] Mark Jones, Karen Grimmer, Ian Edwards, Joy Higgs, and Franziska Trede. 2006. Challenges in applying best evidence to physiotherapy. *Internet Journal of Allied Health Sciences and Practice* 4, 3 (2006), 11.

[28] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature methods* 10, 1 (2013), 64–67.

[29] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1343–1352.

[30] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. (2017).

[31] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

[32] Bongjun Kim and Bryan Pardo. 2018. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–23.

[33] Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*. 2260–2268.

[34] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[35] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. 5–9.

[36] Amanda Kube, Sanmay Das, and Patrick J Fowler. 2019. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 622–629.

[37] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[38] Peter Langhorne, Julie Bernhardt, and Gert Kwakkel. 2011. Stroke rehabilitation. *The Lancet* 377, 9778 (2011), 1693–1702.

[39] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernadino, et al. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on intelligent user interfaces*. ACM, 218–228.

[40] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[41] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. An exploratory study on techniques for quantitative assessment of stroke rehabilitation exercises. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 303–307.

[42] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[43] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[44] Mingkun Li and Ishwar K Sethi. 2006. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence* 28, 8 (2006), 1251–1261.

[45] Andrew F Long, Rosie Kneafsey, and Julia Ryan. 2003. Rehabilitation practice: challenges to effective team working. *International journal of nursing studies* 40, 6 (2003), 663–673.

[46] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.

[47] Susan B O'Sullivan, Thomas J Schmitz, and George Fulk. 2019. *Physical rehabilitation*. FA Davis.

[48] Madhuri Panwar, Dwaipayan Biswas, Harsh Bajaj, Michael Jöbges, Ruth Turk, Koushik Maharatna, and Amit Acharyya. 2019. Rehab-Net: Deep Learning Framework for Arm Movement Classification Using Wearable Sensors for Stroke Rehabilitation. *IEEE Transactions on Biomedical Engineering* 66, 11 (2019), 3026–3037.

[49] Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection* 26, 5 (2020), 584–595.

[50] 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*. ACM, 1135–1144.

[51] Brandon Rohrer, Susan Fasoli, Hermano Igo Krebs, Richard Hughes, Bruce Volpe, Walter R Frontera, Joel Stein, and Neville Hogan. 2002. Movement smoothness changes during stroke recovery. *Journal of Neuroscience* 22, 18 (2002), 8297–8304.

[52] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[53] Saima Safdar, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. 2018. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. *Artificial Intelligence Review* 50, 4 (2018), 597–623.

[54] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 99–109.

[55] Emily Seto, Kevin J Leonard, Joseph A Cafazzo, Jan Barnsley, Caterina Masino, and Heather J Ross. 2012. Developing healthcare rule-based expert systems: case study of a heart failure telemonitoring system. *International journal of medical informatics* 81, 8 (2012), 556–565.

[56] Philip J Smith, Norman D Geddes, and Roger Beatty. 2009. Human-centered design of decision-support systems. In *Human-Computer Interaction*. CRC Press, 263–292.

[57] Katherine J Sullivan, Julie K Tilson, Steven Y Cen, Dorian K Rose, Julie Hershberg, Anita Correa, Joann Gallichio, Molly McLeod, Craig Moore, Samuel S Wu, et al. 2011. Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. *Stroke* 42, 2 (2011), 427–432.

[58] Edward Taub, David M Morris, Jean Crago, Danna Kay King, Mary Bowman, Camille Bryson, Staci Bishop, Sonya Pearson, and Sharon E Shaw. 2011. Wolf motor function test (WMFT) manual. *Birmingham: University of Alabama, CI Therapy Research Group* (2011).

[59] David Webster and Ozkan Celik. 2014. Systematic review of Kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation* 11, 1 (2014), 108.

[60] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 238.

[61] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 694–699.