IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 August 2023, accepted 10 September 2023, date of publication 15 September 2023,
date of current version 21 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316019

RESEARCH ARTICLE

Check It Before You Wreck It: A Guide to STAR-ML for Screening Machine Learning Reporting in Research

RYAN G. L. KOH[1,*], MD ASIF KHAN[2,*], SAJJAD RASHIDIANI[2], SAMAH HASSAN[3], VICTORIA TUCCI[4], THEODORE LIU[2], KARLO NESOVIC[1], DINESH KUMBHARE[1,*], (Member, IEEE), AND THOMAS E. DOYLE[2,5,6,*], (Senior Member, IEEE)

[1]KITE Research Institute, Toronto Rehabilitation Institute, University Health Network (UHN), Toronto, ON M5G 2A2, Canada
[2]Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada
[3]The Institute of Education Research (TIER), UHN, Toronto, ON M5T 1V4, Canada
[4]Faculty of Health Sciences, McMaster University, Hamilton, ON L8S 4L8, Canada
[5]School of Biomedical Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada
[6]Vector Institute for Artificial Intelligence, Toronto, ON M5G 1M1, Canada

Corresponding author: Ryan G. L. Koh (ryan.koh@mail.utoronto.ca)

This work was supported in part by the Canadian Department of National Defence IDEaS under Award CFPMN2-17; and in part by the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada.

*Ryan G. L. Koh and Md Asif Khan contributed equally to this work. Dinesh Kumbhare and Thomas E. Doyle contributed equally to this work.

ABSTRACT Machine learning (ML) is a technique that learns to detect patterns and trends in data. However, the quality of reporting ML in research is often suboptimal, leading to inaccurate conclusions and hindering progress in the field, especially if disseminated in literature reviews that provide researchers with an overview of a field, current knowledge gaps, and future directions. While various tools are available to assess the quality and risk-of-bias of studies, there is currently no generalized tool for assessing the reporting quality of ML in the literature. To address this, this study presents a new screening tool called STAR-ML (Screening Tool for Assessing Reporting of Machine Learning), accompanied by a guide to using it. A pilot scoping review looking at ML in chronic pain was used to investigate the tool. The time it took to screen papers and how the selection of the threshold affected the papers included were explored. The tool provides researchers with a reliable and systematic way to evaluate the quality of reporting of ML studies and to make informed decisions about the inclusion of studies in scoping or systematic reviews. In addition, this study provides recommendations for authors on how to choose the threshold for inclusion and use the tool proficiently. Lastly, the STAR-ML tool can serve as a checklist for researchers seeking to develop or implement ML techniques effectively.

INDEX TERMS Checklist, literature review, machine learning, quality scoring, reporting assessment, research methodology, screening tool.

I. INTRODUCTION

Machine learning (ML) is a rapidly evolving field encompassing a broad range of algorithms designed to perform intelligent predictions based on data [1], [2]. These datasets are often large and complex, typically consisting of millions of unique data points [3], [4]. Recent advances in ML have yielded remarkable progress, with some algorithms achieving a human level of semantic understanding and information extraction, and sometimes the ability to detect abstract patterns with greater accuracy than human experts [5], [6], [7], [8], [9], [10]. By detecting patterns in the data, ML algorithms can extract information, classify data, cluster similar data points, and make predictions for unseen data revealing meaningful insights [1], [11], [12]. These capabilities have numerous applications in fields such as medicine, engineering,

The associate editor coordinating the review of this manuscript and approving it for publication was Mounim A. El Yacoubi.

and finance [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Depending on the amount and type of data and the learning approach, ML algorithms can be categorized into four major groups, i.e., supervised learning (data labels provided), semi-supervised learning (data labels are provided for a small subset of the data, and the rest of the data are unlabeled), unsupervised learning (no data labels provided), and reinforcement learning (develops patterns based on positive and negative rewards) [11], [32].

Supervised learning is a popular and widely used technique for performing classification or regression when the ground truth is known [33], [34]. On the opposite end of the spectrum, in many cases, ground truth labels may not be readily available, or the goal may be to categorize data into separable clusters based on inherent features in the data. In such scenarios, *unsupervised learning* techniques are employed [35], [36]. However, *semi-supervised learning* lies between supervised and unsupervised techniques, where methods take advantage of supervised approaches without needing a complete set of ground truth labels. This method allows leveraging limited labelled data and plentiful unlabeled data [11], [37]. *Reinforcement learning*, on the other hand, is distinct from supervised learning as it does not require labelled input and output pairs. Instead, a learning agent focuses on exploring uncharted territory and exploits known knowledge through trial and error. It is rewarded for desirable behavior and punished for undesirable ones, allowing it to learn and improve over time [11]. Each of these techniques has unique strengths and applications, making them valuable tools for resolving a multitude of challenges across diverse domains.

Based on the data and the objective, a suitable ML or artificial intelligence (AI) algorithm is selected or developed. Due to the powerful nature of ML algorithms, it has become increasingly popular in many different fields. With the increasing number of ML articles in the literature, it becomes useful on occasion to summarize the current knowledge status and gaps in particular fields via a review.

The science of review generally involves gathering research, sifting through it to remove irrelevant or low-quality studies, and summarizing the best evidence that remains. Despite recognizing the need for synthesizing research evidence for over two centuries, it was not until the 20th century that explicit methods for conducting reviews were developed [38]. Review articles help to assess 'what is known' and can aid in assessing what has been studied and what needs to be studied. There are many types of review articles one can use to assess the current state and historical development of the existing literature. By synthesizing the existing research and identifying gaps in knowledge, reviews provide foundations for future research and highlight areas requiring further investigation [38]. Sutton et al. [39] identified 48 review types and categorized them into seven families, whereas Grant and Booth [38] comprised a list of 14 review types, e.g., critical, narrative summary, systematic, meta-analysis, scoping, rapid, and umbrella review.

The objective of a critical review is to demonstrate that the author has conducted a comprehensive examination of the literature and critically evaluated its quality. It entails going beyond a mere description of the literature and encompassing a high degree of analysis and conceptual innovation. Usually, a critical review leads to the development of a hypothesis or model that contributes significantly to the field [38].

Narrative summary reviews help identify and summarize what has been previously published and are great for addressing one or more questions related to a single topic with a broader scope [39]. These can be highly useful in understanding the state of the field depending on the authors' knowledge and experience.

Meta-analysis is a statistical way of combining the findings of similar quantitative studies to produce a more precise effect of the results. By synthesizing and aggregating the data from multiple studies, meta-analysis can provide a more comprehensive understanding of a particular phenomenon or intervention than any individual study could achieve [38], [39], [40].

A scoping review aims to identify the nature and extent of existing research evidence on a particular topic. It typically presents its findings in a tabular format with some narrative commentary. By exploring and defining the boundaries of the topic, it seeks to inform future systematic reviews or primary research. Ultimately, the goal is to provide a comprehensive overview of the available literature and identify knowledge gaps that can guide future research [38], [39].

A systematic review involves systematic literature searching, appraising, and synthesizing research evidence to provide a robust and evidence-based summary of the state of knowledge on a specific topic. The review process often adheres to established guidelines. The aim is to provide a comprehensive analysis of what is known about a particular topic, and identify areas of uncertainty around the findings, including recommendations for practice based on the available evidence. Additionally, similar to a scoping review, the systematic review highlights gaps in knowledge and makes recommendations for future research to fill those gaps [38]. Scoping and systematic reviews are better for collating evidence to identify, select or critically appraise relevant primary research for a specific research question as they follow a standardized protocol.

These reviews typically follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework that provides a minimum set of items for reporting in a review. PRISMA follows a multi-stage process where studies are screened for inclusion/exclusion criteria by at least two reviewers at a title & abstract review stage and a full text review stage before proceeding to any meta-analysis of the studies included in the review [41], [42].

Within the PRISMA 2020 framework, different tools can be used to screen and assess articles [42]. These can be about reporting quality, internal or external validation, and risk-of-bias. These are important factors to consider when collating evidence to answer a research question with the studies in the literature, as it allows comparing articles critically with

confidence [42], [43], [44], [45], [46], [47], [48], [49], [50], [51]. Currently, there are several tools to assess the methodological quality of studies, such as Consolidated Standards of Reporting Trials (CONSORT) [43], Checklist for Critical Appraisal and Data Extraction (CHARMS) [51], Quality Assessment of Diagnostic Accuracy Studies (QUADAS) [49], [50], [52], but there is no generalized tool to assess the quality of ML algorithm reporting in the literature.

Given the increasing usage of ML, quality assessment of articles is essential to maintain high methodological quality and ensure rigorous standards have been followed [53], [54]. Articles that do not report or incorrectly report their algorithm can call into question the quality and transparency of the research. It may also disseminate incorrect or inadequate information that can hinder the progress of scientific knowledge. Thus, it is essential that an evidence-based, validated screening tool be developed to allow for adequate appraisal of the literature for review articles.

Currently, a few tools are associated with ML, such as QUADAS-2 [50], [52], TRIPOD-AI [55], PROBAST-AI [56], STARD-AI [57], SPIRIT-AI [58], CONSORT-AI [59], DECIDE-AI [60], Radiomics Quality Score [61], but these are very specific to applications specifically in the healthcare domain. For example, QUADAS-2 assesses the quality of diagnostic accuracy studies for systematic Reviews. TRIPOD-AI, PROBAST-AI, STARD-AI, SPIRIT-AI, and CONSORT-AI are study design specific, i.e., diagnostic or prognostic model evaluation, risk-of-bias, and reporting guidelines for clinical trials involving AI interventions. DECIDE-AI is a guideline used to report early stage evaluation of AI systems as an intervention in live clinical settings focusing on human factors, clinical utility, and safety [60]. Radiomics Quality Score, on the other hand, is mainly looking at the feature extraction from images. To the best of our knowledge, none of these tools can be used generally and/or applied quickly.

Thus, this paper describes a generalized screening tool (Screening Tool for Assessing Reporting of Machine Learning; STAR-ML) that can assess the **quality of reporting** of ML in research articles quickly and consistently. It also recommends a suitable location for use in the PRISMA framework when reviewing ML-related studies. This paper presents STAR-ML's development, the questions, instructions on use, and its application (pilot) in a scoping review on ML in chronic pain. The tool is intended for quick screening of ML studies but can also be utilized as a guideline when drafting a new manuscript to improve the reporting quality of ML techniques.

II. METHODS

A. DEVELOPMENT OF STAR-ML

The literature was reviewed for existing reporting guidelines and tools for ML studies (e.g., QUADAS-2, CHARMS) [42], [44], [45], [47], [48], [49], [50], [51], [52]. It was identified that two main aspects of ML were significant for any application: *data* and *algorithm*. These aspects were determined through expert consensus as it is important that the data is adequately reported in combination with the rationale and performance of the algorithm. This would enhance reproducibility, improve the use of the algorithms, and encourage knowledge translation. From this perspective, the screening tool was developed with a focus on assessing the implementation of an ML algorithm and identifying if they were reported correctly in the articles.

Through an iterative process, questions in STAR-ML were first developed by expert consensus (RK, AK, DK, TD). These resulted in two versions of STAR-ML with the main changes between the iterations being improvement of the language of the questions and the inclusion of an additional question. These first versions were then piloted in two rounds to assess the functionality of the screening tool [53]. Three independent raters participated in that pilot process.

B. OVERVIEW OF STAR-ML

The set of questions encompassed crucial aspects necessary for reproducing and comprehending the results of the ML algorithm based on the data and parameters used.

1) DATA DOMAIN

The questions in this section pertained to reporting on input features, data quality, and distribution. The origins of data quality issues can be traced back to the early days of computing. High-quality data is crucial for successful ML, as the algorithm's accuracy is dependent on the quality of the data used for training [62], [63]. Providing an algorithm with bad, irrelevant, or faulty data will lead to the ML algorithm reaching an incorrect solution. On the other hand, high-quality data can improve the performance of the ML model and allow for more complex tasks to be solved [64].

2) ALGORITHM DOMAIN

The algorithm section of STAR-ML focuses on the reporting of implementation, training, and performance metrics of the ML algorithm to ensure repeatability. Proper reporting is crucial, as an incorrectly trained or improperly used algorithm can lead to misrepresented results (e.g., higher accuracy due to train-test contamination leading to data leakage) [65], [66], [67], [68], [69].

C. STAR-ML QUESTIONS

Table 1 presents the list of questions in STAR-ML with a summary explanation of each question's focus.

1) DATA DOMAIN

Question 1: *"Did the study report the data used?"* Focuses on whether the input features of the algorithm are known. It is important to know exactly what is used in the algorithm to be able to understand and reproduce results seen in a particular study.

Question 2: *"Did the study report data quality or data pre-processing?"* Focuses on data pre-processing and whether

steps were taken to address issues with the data, such as handling missing values, duplicate entries, incorrect, corrupted, or incorrectly formatted entries. Proper data pre-processing is essential to ensure the validity and reliability of study results [70], [71].

Question 3: *"Did the study report data distribution and if imbalanced did they handle it?"* Focuses on the reporting of the distribution of data and whether any steps were taken to address imbalances (if any). ML models trained with imbalanced data can lead to incorrect insights and overestimated generalizable performance, making it vital for taking into consideration [72], [73], [74].

Question 4: *"Did the study report data normalization?"* Focuses on whether the features of the data were transformed into some standard range or space. Though data normalization is considered one of the pre-processing steps, attention was given with a dedicated question given its importance to model performance. The absence of feature normalization can result in issues where certain features carry more weight than others, particularly in algorithms that utilize distance-based metrics such as k-nearest neighbors [75], [76].

2) ALGORITHM DOMAIN

Question 5: *"Did the study report any rationale behind their choice of algorithm?"* Focuses on whether there was any rationale for the choice of ML algorithm. Except for the cases of novel algorithms being developed and described in a study, different ML techniques have strengths and weaknesses depending on the data and problem. For example, support vector machines (SVM) are designed to provide the best boundaries between classes even when there is a limited amount of data. However, SVMs are less practical in multi-class problems as the number of decision boundaries increases with the number of classes [77], [78], [79]. A convolutional neural network (CNN), on the other hand, can take image or video data as input. However, since a CNN is more complex, it may require more training examples to be able to provide an adequate solution [80], [81]. Thus, the rationale behind the choice of algorithm is crucial, as different ML algorithms are designed for different purposes.

Question 6: *"Focusing on modeling, did the study report any measure/s to address their model bias?"* Bias can occur during data collection, data pre-processing, data selection/splitting, model training, and model evaluation. The bias observed in a resulting or final AI/ML model is model bias or algorithmic bias [82], [83], [84], [85]. Several sources of bias can occur throughout the pipeline, from the data collection to the developed ML model. This can affect the performance of the model and the conclusions that the model learns [82], [86]. For example, bias in the data collection can lead to learned features in only certain conditions; if all examples come from females, it may draw incorrect conclusions when observing an example that is male. Thus, it is important to ensure that biases are "kept in check".

Question 7: *"Focusing on reproducibility, did the study report their model parameter/s?"* Focuses on whether the parameters of the ML model(s) are reported. Model parameters are important to be able to reproduce the results reported in a particular study. Without the parameters, it is extremely difficult to nearly impossible to train a similar model depending on the ML algorithm [87]. For example, if a CNN was used to classify images of cats vs. dogs, deciding the parameters to reproduce the results would be challenging. A CNN used for image classification requires many tunable parameters, such as architecture (i.e., number of layers, number of neurons), activation function, loss function, and the number of filters, among other parameters [87], [88].

Question 8: *"Focusing on model training, did the study report validation technique(s) for the model?"* Model validation in ML is the process of evaluating how well a trained model performs on unseen data (e.g., test set), in order to ensure that the model is able to generalize to new data. Model validation helps to assess the reliability of the model's predictions and to identify any issues such as overfitting, underfitting, or bias that may impact the model's performance [89], [90], [91], [92]. There are several techniques for model validation in ML, including holdout validation, cross-validation, leave-one-out validation, and bootstrapping [93]. By using these techniques, ML models can be assessed in terms of performance, and informed decisions can be made about improving them.

Question 9: *"Focusing on model test/validation, did the study report any performance evaluation metric of the used algorithm?"* Focuses on how ML model(s) performed on the task based on a new example and/or dataset. The model's effectiveness can be understood by assessing the measures like accuracy, precision, recall, F1 score, and other relevant performance metrics of the model(s) [94]. It is also crucial for determining the suitability of the model(s) for the intended application and making informed decisions regarding improving performance [95].

D. INSTRUCTIONS FOR USE

1) SCORING

Each question can be answered with either a Yes (1 point) or No (0 points).

Question 1: A score of 1 should be given if the data features were reported in the study. Otherwise, a score of 0 should be assigned.

Question 2: A score of 1 should be given if any data cleaning or pre-processing was reported in the study. Otherwise, a score of 0 should be assigned.

Question 3: A score of 1 should be given if the data distribution is reported (e.g., descriptive statistics) and information is provided regarding data imbalance or if there were data imbalances, but it was reported that techniques were used to address that. Otherwise, a score of 0 should be assigned.

Question 4: A score of 1 should be given if the input features were transformed into a standard or normalized range. However, in special cases (i.e., all features have the same unit,

TABLE 1. Screening tool for assessing reporting of machine learning (STAR-ML): An overview of the questions and the explanations for each question.

Domain	Questions	Explanation
Data		
$Q1$	Did the study report the data used?	Which or what data were used in the study while describing data features.
$Q2$	Did the study report data quality or data pre-processing?	To assess the study's validity and reliability of the data analysis. It also addresses if the missing data handling and outlier removal steps were considered by the authors.
$Q3$	Did the study report data distribution and if imbalanced did they handle it?	The data distribution should be reported with descriptive statistics and how the authors handled the data imbalance if there was any.
$Q4$	Did the study report data normalization?	Data or feature normalization is a necessary step from which most of the AI/ML algorithms can benefit, including the model's numerical stability.
Algorithm		
$Q5$	Did the study report any rationale behind their choice of algorithm?	Except in the case of tailored/novel AI/ML algorithms, the rationale for which the particular algorithm was used should be stated.
$Q6$	Focusing on modeling, did the study report any measure/s to address their model bias?	ML model biases should be addressed to get reliable performance.
$Q7$	Focusing on reproducibility, did the study report their model parameter/s?	Model parameters or hyperparameters should be reported for reproducibility of the result as most of the ML algorithms have tunable parameters.
$Q8$	Focusing on model training, did the study report validation technique(s) for the model?	Model validation is an integral part of any ML model and is expected to be reported.
$Q9$	Focusing on model test/validation, did the study report any performance evaluation metric of the used algorithm?	Different algorithms have different evaluation metrics based on their fundamental design and it is a major way to evaluate the performance of the model on the particular data.

the reason why the data were not standardized or normalized is mentioned), a score of 1 should be assigned. Otherwise, a score of 0 should be assigned.

Question 5: A score of 1 should be given if a rationale was provided for the ML algorithm used in the study. Otherwise, a score of 0 should be assigned.

Question 6: A score of 1 should be given if model bias handling was reported in the study. Otherwise, a score of 0 should be assigned.

Question 7: A score of 1 should be given if parameters needed to build the ML models were reported. Otherwise, a score of 0 should be assigned.

Question 8: A score of 1 should be given if ML model training and validation techniques were reported in the study. Otherwise, a score of 0 should be assigned.

Question 9: A score of 1 should be given if performance evaluation metrics were reported in the study. Otherwise, a score of 0 should be assigned.

2) ASSESSING A SPECIFIC STUDY USING STAR-ML

The PRISMA framework [41] is a widely used tool to help authors improve the reporting of systematic or scoping reviews. While using the PRISMA 2020 flow diagram [42] template (e.g., systematic reviews), STAR-ML can be used for screening at the "Reports assessed for eligibility" stage to only include well-reported research into the full text review.

However, STAR-ML can be used independently or in conjunction with other frameworks to assess the reporting quality of a study that used ML.

E. PILOT LITERATURE REVIEW

STAR-ML was piloted in a scoping review on the topic of ML in chronic pain.

1) SEARCH STRATEGY

Search terms were developed in Ovid MEDLINE and then adapted for other databases. The search string was developed and finalized after multiple reviews and iterations with subject-matter experts and the assistance of a health science librarian specialized in conducting literature searches. The searches were carried out across 4 electronic databases from 2012-2022:

1) MEDLINE
2) Web of Science Core Collection
3) ACM Digital Library
4) IEEE Xplore

The final search was conducted on February 28, 2022. The final search string for Ovid MEDLINE can be found below:

1) artificial intelligence/ or exp machine learning/ or natural language processing/ or neural networks, computer/ or cluster analysis/
2) artificial* intelligen*.ti,ab,kf,kw.
3) machine learning.ti,ab,kf,kw.
4) (deep learning or convolutional neural network or artificial neural network).ti,ab,kf,kw.
5) (cluster analysis or (unsupervised adj2 learning)).ti,ab,kf,kw.
6) natural language processing.ti,ab,kf,kw.
7) computer neural network*.ti,ab,kf,kw.
8) or/1-7
9) Chronic Pain/
10) ((Chronic* or Recurrent or Persistent*) adj3 pain*).ti,ab,kf,kw.
11) or/9-10
12) 8 and 11
13) limit 12 to english language
14) not (animals/ not (humans/ and animals/))
15) limit 14 to journal article
16) limit 15 to ("review articles" or meta analysis or "systematic review" or comment or editorial)
17) 15 not 16
18) exp Neoplasms/
19) 17 not 18
20) remove duplicates from 19
21) limit 20 to yr="2012 -Current"

2) INCLUSION & EXCLUSION CRITERIA

Studies were included if they satisfied the following criteria:

1) Studies published in English
2) Studies involving only human participants/data
3) Original research article, i.e., not a review article or letter
4) Peer-reviewed
5) Studies focused on chronic pain
6) Used ML methods
7) Studies focused on physically adults (17+)
8) Studies excluding only healthy participants or synthetic data
9) Studies scored 6 or more in STAR-ML

F. EVALUATION

In this scoping review, STAR-ML was used as one of the exclusion criteria at the "Reports assessed for eligibility stage" during full text review. In total, 4 raters, i.e., 2 experienced and 2 less experienced ML raters scored articles and articles meeting the threshold for inclusion were included for data extraction (after applying all other exclusion criteria).

Out of the 289 studies considered in the full text screening phase, 111 were excluded based on the other pre-defined exclusion criteria. The remaining 178 studies were divided among four raters and screened using STAR-ML. As STAR-ML achieved high inter-rater reliability (0.93, lower bound: 0.83, upper bound: 0.97), which was computed based on a mean-rating ($n = 3$), absolute-agreement, 2-way mixed-effects model for the total scores between 3 raters in the previous study [53], a single rater was used to score the assigned set of articles. Prior to working on the articles in the scoping review, the raters underwent training on a set of 10 articles to align the level of understanding.

Initially, experts agreed that during full text review, a study should be included if the STAR-ML score is more than 50%, which corresponded to $a \geq 5$ out 9 [53]. A score of 6 or more was chosen for the pilot review to ensure that included studies were better than the minimal score determined initially by the experts.

After the full text screening, the studies that were included for data extraction were then scored by a second rater whose experience depended on the experience of the initial rater of the included study (i.e., if the initial rater had ML experience, then the second rater was less experienced, and vice-versa). The group of raters was the same in the two steps. This was to understand better how the scoring would be affected by varying ML experiences. These raters were blinded to whether the studies were included or not.

In addition, the quality of the reporting of the ML techniques was also examined. One experienced rater and an independent expert (i.e., not one of the four raters) assessed the quality of each included study on the reproducibility and the correctness (i.e., rigorous methodology) of the described ML techniques and procedures.

A study was deemed reproducible if, with the given information, one could likely reproduce the pipeline and expect to get a similar result; it was deemed not reproducible if there was missing information that would be required to reproduce the procedure, e.g., the exact input features to the ML algorithm was not reported, no architecture reported for neural network based models, no performance metrics to compare results, etc. Similarly, a study was deemed correct if the study used appropriate procedures given the data; otherwise, it was deemed incorrect. For instance, if a study had imbalanced data and the study only used classification accuracy as a performance metric, this would overestimate the performance, and the study would be deemed not correct (i.e., implementation procedures were not adequate). While STAR-ML was primarily designed to assess the quality of reporting of ML studies, being able to correlate the score with the quality of the work would allow for a more informed decision on setting the threshold for including studies based on quantitative data. This would also provide examples of the types of ML studies that could be included at different threshold levels, helping to guide researchers in selecting appropriate studies for their works.

Lastly, the time required to use STAR-ML to screen studies was evaluated for raters with varying levels of experience in ML. Raters 1 and 3 were experienced in ML, while Raters 2 and 4 had basic experience. This provided insights into the feasibility and efficiency of using STAR-ML as a screening tool for ML studies in terms of helping to make informed decisions on allocating resources for conducting reviews or for independent use.

III. RESULTS

Fig. 1 presents the distribution of STAR-ML scores for all studies included in the scoping review, as well as the scores assigned by pairs of raters (i.e., Raters 1 and 2, and Raters 3 and 4) for subsets of the studies. The mean and standard deviation of STAR-ML scores for all studies were 5.640 ± 2.054. The mean and standard deviation of scores for each rater on the subsets of the included studies were 7.909 ± 0.707, 8.091 ± 1.071, 7.519 ± 0.935, and 7.667 ± 1.494, respectively.

Table 2 shows the number of studies that would have been included at each score level from 6 to 9, along with the number of studies that were deemed reproducible and/or done correctly by experts on our team. As expected, the number of studies to be included decreases as the score threshold increases. Interestingly, the percentage of reproducible and rigorous studies peaks at a score of 8. Table 3 provides the number of studies for each rater that would have been included at each score level from 6 to 9, along with the number of studies that were deemed reproducible and/or done correctly by the experts on our team. These results are also visualized in Fig. 2 and Fig. 3.

Fig. 4 shows a comparison of the average time it took to screen a study in this work and in [53]. The mean and standard deviation of the average time was 4.733 ± 2.101 minutes and 4.701 ± 0.644 minutes for the former study [53] and the current pilot, respectively. The average time of each rater was

TABLE 2. Number of included studies by STAR-ML thresholds that were deemed reproducible, implemented correctly or both (reproducible and implemented correctly).

STAR-ML Score	Number of Studies			
Threshold	Included in Review	Reproducible	Correctness	Both
6	60	42 (70.00%)	46 (76.67%)	34 (56.67%)
7	37	27 (72.97%)	30 (81.08%)	23 (62.16%)
8	19	15 (78.95%)	17 (89.47%)	13 (68.42%)
9	5	3 (60.00%)	5 (100.00%)	3 (60.00%)

TABLE 3. Number of included studies by STAR-ML thresholds and Raters that were deemed reproducible, implemented correctly or both (reproducible and implemented correctly).

Rater 1				
STAR-ML Score	**Number of Studies**			
Threshold	Included in Review	Reproducible	Correctness	Both
6	10	9 (90.00%)	6 (60.00%)	6 (60.00%)
7	4	4 (100.00%)	3 (75.00%)	3 (75.00%)
8	1	1 (100.00%)	0 (0.00%)	0 (0.00%)
9	0	0 (NA)	0 (NA)	0 (NA)
Rater 2				
STAR-ML Score	**Number of Studies**			
Threshold	Included in Review	Reproducible	Correctness	Both
6	23	19 (82.61%)	17 (73.91%)	15 (65.22%)
7	18	14 (77.78%)	13 (72.22%)	11 (61.11%)
8	11	10 (90.91%)	10 (90.91%)	9 (81.82%)
9	2	2 (100.00%)	2 (100.00%)	2 (100.00%)
Rater 3				
STAR-ML Score	**Number of Studies**			
Threshold	Included in Review	Reproducible	Correctness	Both
6	10	3 (30.00%)	7 (70.00%)	3 (30.00%)
7	2	1 (50.00%)	2 (100.00%)	1 (50.00%)
8	1	1 (100.00%)	1 (100.00%)	1 (100.00%)
9	0	0 (NA)	0 (NA)	0 (NA)
Rater 4				
STAR-ML Score	**Number of Studies**			
Threshold	Included in Review	Reproducible	Correctness	Both
6	17	11 (64.71%)	16 (94.12%)	10 (58.82%)
7	12	8 (66.67%)	12 (100.00%)	8 (66.67%)
8	6	3 (50.00%)	6 (100.00%)	3 (50.00%)
9	3	1 (33.33%)	3 (100.00%)	1 (33.33%)

4.932 ± 1.934, 5.500 ± 1.487, 4.085 ± 0.905, and 4.286 ± 1.235 minutes in the current pilot.

IV. DISCUSSION

This paper presents the development and instructions for using a novel screening tool for assessing the reporting quality of ML studies, called STAR-ML. Additionally, the study provides valuable insight into determining the appropriate threshold for inclusion, along with the expected quality of ML studies that may be observed at each level.

As expected, setting a higher threshold for inclusion leads to a decrease in the number of studies meeting the criterion. Notably, Table 2 shows that the number of studies deemed to be done correctly (based on expert opinion) increases as the threshold increases, while reproducibility follows the same trend until a score of 9. An 'elbow' point is observed at a score of 8 out of 9. However, further investigation in Table 3 reveals that the largest percentage of studies' ML techniques are identified as done correctly in each rater's pool of studies when a threshold of 7 out of 9 is used, except notably in Rater 2's pool.

This would suggest that raters with a good understanding of ML could use a threshold of 7 out of 9 for inclusion, as it will result in studies that are mostly deemed to be done correctly. However, if raters have different levels of experience with ML, then a threshold of 8 out of 9 should be used as this correlated with almost every study's ML techniques being identified as being implemented and used appropriately. It is important to note that the appropriate threshold for inclusion may vary depending on the specific context and research question.

In addition, the finding suggests that STAR-ML scores seem to associate well with algorithm implementation, correct use, and reproducibility of the described algorithms by the authors (i.e., a higher percentage of included studies increased as STAR-ML score increases) with the exception of the percentage of included studies for the 'Reproducible' category at STAR-ML score of 9. Arguably, correct implementation and use are of more importance in collecting evidence for a scoping or systematic review, as the findings of a review must be based on results from correctly implemented approaches.

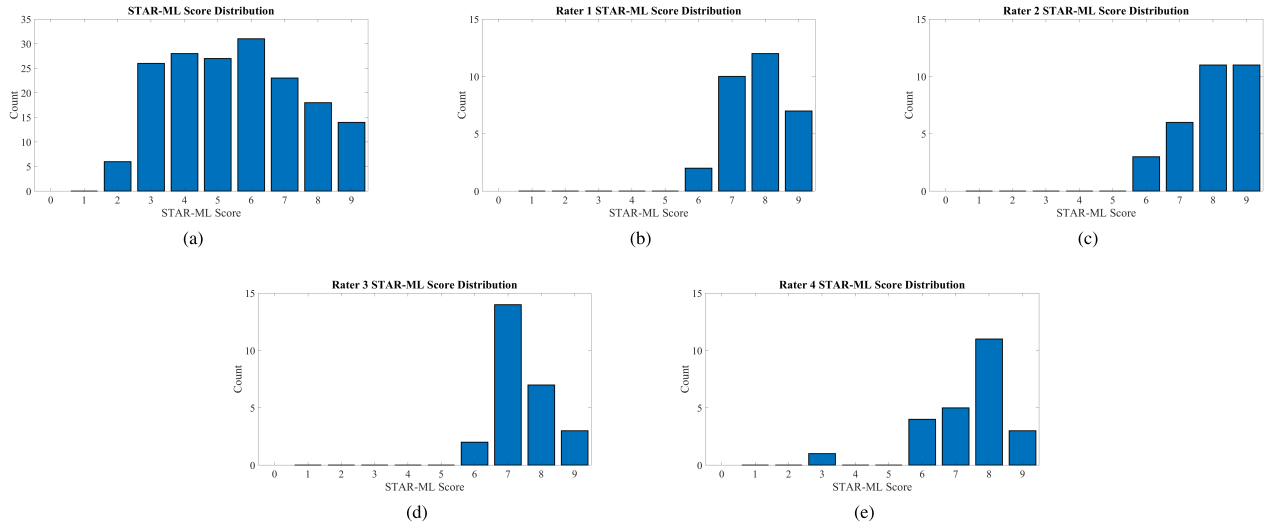

FIGURE 1. a) Distribution of the STAR-ML scores of all screened studies during full text review. b, c, d, e) STAR-ML score distribution of each rater on their subset of studies included for data extraction. Raters 1 & 2 and 3 & 4 reviewed the same subset of studies. Mean and standard deviations of each rater: Rater 1: 7.909 ± 0.707; Rater 2: 8.091 ± 1.071; Rater 3: 7.519 ± 0.935; Rater 4: 7.667 ± 1.494.

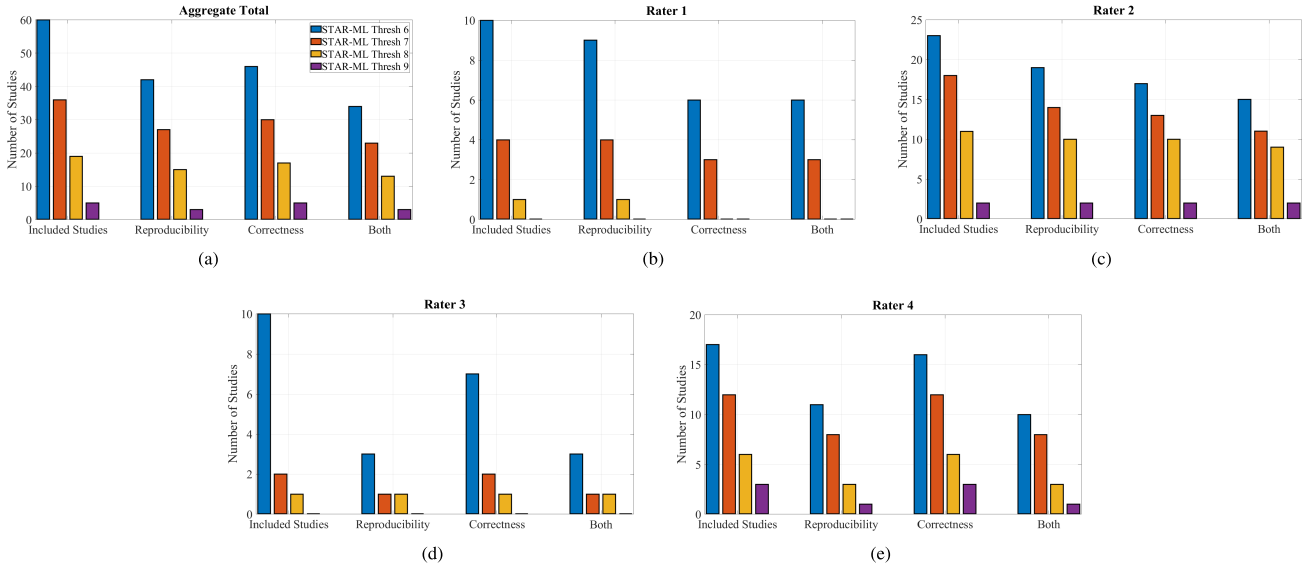

FIGURE 2. a) Shows the aggregated number of studies in terms of included studies, reproducibility, correctness, and Both (Reproducible and Correct). b, c, d, e) Shows the number of studies by each rater in terms of Included Studies, reproducibility, correctness, and Both (Reproducible and Correct).

It is important to note that the reproducibility and correctness score of each study was reached through a consensus and discussion among experts on our team for each study included in the data extraction of the review. Although the results suggest that the STAR-ML score is associated with reproducibility and correctness (i.e., a higher STAR-ML score leads to a higher likelihood of reproducibility and correctness), this analysis was not performed using a validated critical appraisal tool for ML, as there is no such tool to our knowledge. Therefore, further investigation and/or development of such a tool will be necessary to better assess the validity of the ML techniques used for their reproducibility and appropriate use/implementation. Nonetheless, the analysis explained here suggests that a higher STAR-ML score generally indicates a higher reproducibility and appropriate use of ML techniques.

It was also observed that novice ML raters tend to score papers higher on average than experienced raters, i.e., Rater 1: 7.909 vs. Rater 2: 8.091; Rater 3: 7.519 vs. Rater 4: 7.667. This could be due to novice ML raters being more cautious and giving the benefit of the doubt to the papers when they are unsure if a study meets certain criteria. Combining this information with what was observed in Table 2, this provides a strong case for setting the threshold at 8 or above. This can balance the trade-off between the number of papers included, the quality of the paper and the differences in ML experience

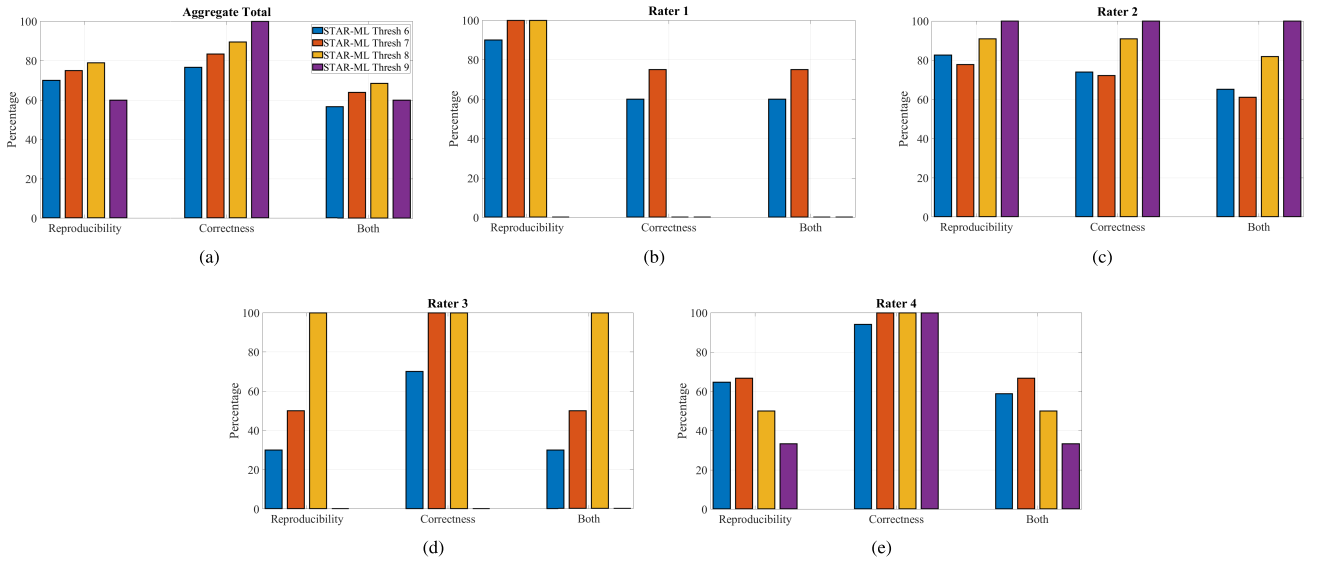

FIGURE 3. a) Shows the aggregate total of the percentage of Included Studies in terms of Reproducibility, Correctness, and Both (Reproducible and Correct) from all raters. b, c, d, e) Shows the percentage of the Included Studies for each rater in terms of Reproducibility, Correctness, and Both (Reproducible and Correct). Note that Rater 1 and 3 did not have any included studies at a threshold of 9.

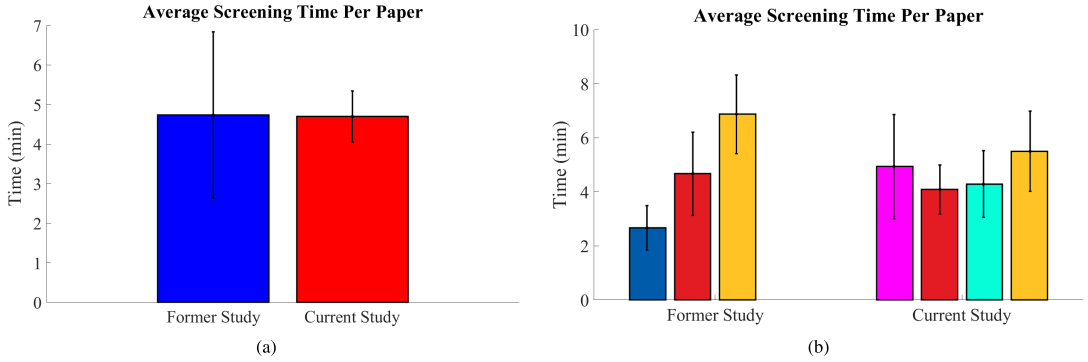

FIGURE 4. a) The average time taken by the raters to rate articles using STAR-ML. b) Average times of each rater. The left bars represent timing information from the former study [53] & the right bars represent the current pilot scoping review results. Note: Raters with the same colour are indicating the same rater in both studies.

of the raters. Thus, we recommend using a threshold of 8 to balance the number of papers included, maintain a high likelihood of high-quality papers, and account for potential differences in the raters' experience with ML. However, this threshold can be adjusted based on the researcher's objective and research question.

The time required to screen a paper using STAR-ML was consistent with what was observed in the previous study [53]. On average, raters tended to screen papers within 4-5 minutes, indicating the tool's feasibility for rapid screening of studies that use ML, especially in a review. Two raters, novices in ML, from the previous work [53] also screened papers in this pilot and showed consistently faster screening time, closer to the observed average time of 4.70 minutes (Fig. 4b: red and yellow bars). This suggests that with an initial 'orientation' of the tool, even novice raters can use the tool to screen studies as quickly as more experienced raters. An additional empirical observation was made that raters might initially be more lenient with their scoring as they become familiar with the tool, and after calibration, are better able to understand the full spectrum of the score, e.g., what types of studies would score low versus what types of studies would score high. Thus, it is recommended to have an initial calibration period with five to ten articles to familiarize raters with the tool.

As noted above, STAR-ML is a tool mainly to assess the quality of reporting of ML algorithm in an article. STAR-ML will best be used in combination with a validated ML quality assessment tool. STAR-ML would screen papers and include those that are likely of high quality in terms of reproducibility and correctness, where a quality assessment tool can then determine if they are indeed of high quality. This would allow for rapid screening of articles as STAR-ML would take a much shorter time than directly assessing the quality of every single article.

STAR-ML can also be used by researchers to self-assess their research manuscripts and improve the reporting of their

work related to ML. Additionally, it can act as a guide or checklist for researchers developing or applying ML techniques. The comprehensive set of criteria included in the tool ensures that researchers consider all relevant factors and best practices when working with or developing ML algorithms. As a result, using the tool can lead to more accurate and robust ML models, improving the quality and reliability of research findings. Thus, STAR-ML can play a significant role in facilitating best practices in the field of ML, accelerating progress and advancing knowledge in the field.

Further work is currently underway by the authors to analyze the tool in various domains, as the study was only piloted in the chronic pain domain. Additionally, work is in progress to assess the generalizability of the tool in terms of ML users from different educational backgrounds and geographic regions. The current version (version 2) of STAR-ML is being published to enable researchers to only include high-quality ML research papers in scoping and systematic reviews to draw accurate conclusions and disseminate high-quality knowledge. Additionally, ongoing efforts will focus on validating the findings presented here and improving the tool to better meet the needs of researchers in the field.

V. CONCLUSION

This paper presents the development of a new generalized tool for assessing the reporting quality of ML in studies that can be used for screening studies during full text screening in reviews. The instructions on using the tool and selecting the threshold for inclusion have been investigated. Depending on the researcher's objective and research question, the results of this study suggest that a score of 7 or 8 on the STAR-ML tool will increase the likelihood that studies included in reviews are of high quality in terms of correctness and reproducibility. The average time required to screen a study using the tool is about $4-5$ minutes, and it is best used by researchers with scientific and ML experience. An initial calibration period to familiarize with the tool before use is also recommended. Additionally, the tool can help new researchers improve the reporting of ML in their manuscripts. Furthermore, the tool can serve as a practical checklist for researchers seeking to develop or implement machine learning techniques effectively, thus promoting best practices in the field and improving the quality and reliability of research findings.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[2] A. Panesar, *Machine Learning and AI for Healthcare*. Cham, Switzerland: Springer, 2019.

[3] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" 2015, *arXiv:1511.06348*.

[4] F. van Wyk, A. Khojandi, R. Kamaleswaran, O. Akbilgic, S. Nemati, and R. L. Davis, "How much data should we collect? A case study in sepsis detection using deep learning," in *Proc. IEEE Healthcare Innov. Point Care Technol. (HI-POCT)*, Nov. 2017, pp. 109–112.

[5] J. Shen, C. J. P. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S. Y. Wong, P.-H. Fang, and W.-K. Ming, "Artificial intelligence versus clinicians in disease diagnosis: Systematic review," *JMIR Med. Informat.*, vol. 7, no. 3, Aug. 2019, Art. no. e10010.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[7] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, Oct. 2018.

[8] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 29–38.

[9] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated Machine Learning: Methods, Systems, Challenges*. Cham, Switzerland: Springer, 2019.

[10] K. Siau and W. Wang, "Building trust in artificial intelligence, machine learning, and robotics," *Cutter Bus. Technol. J.*, vol. 31, no. 2, pp. 47–53, 2018.

[11] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA, USA: O'Reilly Media, 2019, p. 851.

[12] T. M. Mitchell, *Machine Learning*, vol. 1. New York, NY, USA: McGraw-Hill, 2007.

[13] S. Bayat, G. M. Babulal, S. E. Schindler, A. M. Fagan, J. C. Morris, A. Mihailidis, and C. M. Roe, "GPS driving: A digital biomarker for preclinical Alzheimer disease," *Alzheimer's Res. Therapy*, vol. 13, no. 1, pp. 1–9, Dec. 2021.

[14] Q. A. Rahman, T. Janmohamed, M. Pirbaglou, H. Clarke, P. Ritvo, J. M. Heffernan, and J. Katz, "Defining and predicting pain volatility in users of the manage my pain app: Analysis using data mining and machine learning methods," *J. Med. Internet Res.*, vol. 20, no. 11, Nov. 2018, Art. no. e12001.

[15] K. Nesovic, R. G. L. Koh, A. A. Sereshki, F. S. Zadeh, M. R. Popovic, and D. Kumbhare, "Ultrasound image quality evaluation using a structural similarity based autoencoder," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 4002–4005.

[16] D. Jiménez-Grande, S. F. Atashzar, V. Devecchi, E. Martinez-Valdes, and D. Falla, "A machine learning approach for the identification of kinematic biomarkers of chronic neck pain during single- and dual-task gait," *Gait Posture*, vol. 96, pp. 81–86, Jul. 2022.

[17] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and natural language processing of social media data for event detection in smart cities," *Sustain. Cities Soc.*, vol. 85, Oct. 2022, Art. no. 104026.

[18] M. Dousty, D. J. Fleet, and J. Zariffa, "Hand grasp classification in egocentric video after cervical spinal cord injury," *IEEE J. Biomed. Health Informat.*, early access, Apr. 24, 2023, doi: 10.1109/JBHI.2023.3269692.

[19] N. A. Diptu, M. A. Khan, S. Debnath, A. A. Imam, A. M. H. Rakib, K. A. A. Ador, and R. M. Rahman, "Early detection of glaucoma using fuzzy logic in Bangladesh context," in *Proc. Int. Conf. Intell. Syst. (IS)*, Sep. 2018, pp. 87–93.

[20] B. Shi, A. Tay, W. L. Au, D. M. L. Tan, N. S. Y. Chia, and S.-C. Yen, "Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 7, pp. 2256–2267, Jul. 2022.

[21] R. G. L. Koh, M. Balas, A. I. Nachman, and J. Zariffa, "Selective peripheral nerve recordings from nerve cuff electrodes using convolutional neural networks," *J. Neural Eng.*, vol. 17, no. 1, Jan. 2020, Art. no. 016042.

[22] B. Khailany, "Accelerating chip design with machine learning," in *Proc. ACM/IEEE 2nd Workshop Mach. Learn. CAD (MLCAD)*, Nov. 2020, p. 33.

[23] M. Wadi, "Fault detection in power grids based on improved supervised machine learning binary classification," *J. Electr. Eng.*, vol. 72, no. 5, pp. 315–322, Sep. 2021.

[24] S. J. Park, B. Bae, J. Kim, and M. Swaminathan, "Application of machine learning for optimization of 3-D integrated circuits and systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 6, pp. 1856–1865, Jun. 2017.

[25] Y. Reich and S. V. Barai, "Evaluating machine learning models for engineering problems," *Artif. Intell. Eng.*, vol. 13, no. 3, pp. 257–272, Jul. 1999.

[26] D. Zhang, "Machine learning and software engineering," *Softw. Quality J.*, vol. 11, no. 3, pp. 87–119, 2003.

[27] Y. Reich, "Machine learning techniques for civil engineering problems," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 12, no. 4, pp. 295–310, Jul. 1997.

[28] M. F. Dixon, I. Halperin, and P. Bilokon, *Machine Learning in Finance*, vol. 1170. Cham, Switzerland: Springer, 2020.

[29] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," *J. Behav. Experim. Finance*, vol. 32, Dec. 2021, Art. no. 100577.

[30] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: An application to default risk analysis," *SSRN Electron. J.*, vol. 19, pp. 1–44, Aug. 2019.

[31] L. Gan, H. Wang, and Z. Yang, "Machine learning solutions to challenges in finance: An application to the pricing of financial products," *Technol. Forecasting Social Change*, vol. 153, Apr. 2020, Art. no. 119928.

[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[33] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised machine learning: A brief primer," *Behav. Therapy*, vol. 51, no. 5, pp. 675–687, Sep. 2020.

[34] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, Jun. 2007.

[35] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.

[36] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: A review," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 120–154, 2010.

[37] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 109, no. 2, pp. 373–440, Apr. 2020.

[38] M. J. Grant and A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," *Health Inf. Libraries J.*, vol. 26, no. 2, pp. 91–108, Jun. 2009.

[39] A. Sutton, M. Clowes, L. Preston, and A. Booth, "Meeting the review family: Exploring review types and associated information retrieval requirements," *Health Inf. Libraries J.*, vol. 36, no. 3, pp. 202–222, Sep. 2019.

[40] H. Leary and A. Walker, "Meta-analysis and meta-synthesis methodologies: Rigorously piecing together research," *TechTrends*, vol. 62, no. 5, pp. 525–534, Sep. 2018.

[41] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gotzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration," *BMJ*, vol. 339, no. 1, p. b2700, Dec. 2009.

[42] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Int. J. Surg.*, vol. 88, Apr. 2021, Art. no. 105906.

[43] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman, "CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials," *Int. J. Surg.*, vol. 10, no. 1 pp. 1–12, 2011.

[44] W. Sauerbrei, S. E. Taube, L. M. McShane, M. M. Cavenagh, and D. G. Altman, "Reporting recommendations for tumor marker prognostic studies (REMARK): An abridged explanation and elaboration," *J. Nat. Cancer Inst.*, vol. 110, no. 8, pp. 803–811, Aug. 2018.

[45] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, D. Moher, D. Rennie, H. C. W. de Vet, and J. G. Lijmer, "The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration," *Ann. Internal Med.*, vol. 138, no. 1, p. W1, Jan. 2003.

[46] E. V. Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, "Strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies," *Lancet*, vol. 335, no. 7624, pp. 806–808, Oct. 2007.

[47] A. C. J. Janssens, J. P. Ioannidis, C. M. van Duijn, J. Little, and M. J. Khoury, "Strengthening the reporting of genetic risk prediction studies: The grips statement," *PLOS Med.*, vol. 8, Mar. 2011, Art. no. e1000420.

[48] J. P. T. Higgins, D. G. Altman, P. C. Gotzsche, P. Juni, D. Moher, A. D. Oxman, J. Savovic, K. F. Schulz, L. Weeks, and J. A. C. Sterne, "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials," *BMJ*, vol. 343, no. 2, p. d5928, Oct. 2011.

[49] P. Whiting, A. W. Rutjes, J. B. Reitsma, P. M. Bossuyt, and J. Kleijnen, "The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews," *BMC Med. Res. Methodol.*, vol. 3, no. 1, pp. 1–13, Dec. 2003.

[50] P. F. Whiting, A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. Sterne, and P. M. Bossuyt, "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," *Ann. Internal Med.*, vol. 155, no. 8, pp. 529–536, 2011.

[51] K. G. M. Moons, J. A. H. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J. B. Reitsma, and G. S. Collins, "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist," *PLoS Med.*, vol. 11, no. 10, Oct. 2014, Art. no. e1001744.

[52] S. Jayakumar, V. Sounderajah, P. Normahani, L. Harling, S. R. Markar, H. Ashrafian, and A. Darzi, "Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: A meta-research study," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–13, Jan. 2022.

[53] M. A. Khan, R. G. L. Koh, S. Hassan, T. Liu, V. Tucci, D. Kumbhare, and T. E. Doyle, "STAR-ML: A rapid screening tool for assessing reporting of machine learning in research," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, Sep. 2022, pp. 336–341.

[54] C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review," *BMJ*, vol. 2021, p. n2281, Oct. 2021.

[55] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *Ann. Internal Med.*, vol. 162, no. 1, pp. 55–63, Jan. 2015.

[56] G. S. Collins, P. Dhiman, C. L. Andaur Navarro, J. Ma, L. Hooft, J. B. Reitsma, P. Logullo, A. L. Beam, L. Peng, B. Van Calster, M. van Smeden, R. D. Riley, and K. G. Moons, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence," *BMJ Open*, vol. 11, no. 7, Jul. 2021, Art. no. e048008.

[57] V. Sounderajah et al., "Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol," *BMJ Open*, vol. 11, no. 6, Jun. 2021, Art. no. e047709.

[58] S. C. Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension," *Lancet Digit. Health*, vol. 2, no. 10, p. m3210, Sep. 2020.

[59] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G. S. Collins, and A. D. J. Deeks, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," *Lancet Digital Health*, vol. 2, no. 10, pp. e537–e548, 2020.

[60] B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nature Med.*, vol. 28, no. 5, pp. 924–933, 2022.

[61] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, Dec. 2017.

[62] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. S. Mittal, and V. Munigala, "Overview and importance of data quality for machine learning tasks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3561–3562.

[63] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *Int. J. Adv. Softw.*, vol. 10, no. 1, pp. 1–20, 2017.

[64] R. Franklin. (Apr. 2020). *How Data Quality Impacts Machine Learning—Precisely*. Accessed: Sep. 3, 2023. [Online]. Available: https://www.precisely.com/blog/data-quality/data-quality-impact-machine-learning

[65] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ML-based science," 2022, *arXiv:2207.07048*.

[66] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[67] E. Yagis, S. W. Atnafu, A. G. Seco de Herrera, C. Marzi, R. Scheda, M. Giannelli, C. Tessa, L. Citi, and S. Diciotti, "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Sci. Rep.*, vol. 11, no. 1, p. 22544, Nov. 2021.
[68] Y. Ji, A. Sun, J. Zhang, and C. Li, "A critical study on data leakage in recommender system offline evaluation," *ACM Trans. Inf. Syst.*, vol. 41, no. 3, pp. 1–27, Jul. 2023.
[69] S. Kaufman, S. Rosset, and C. Perlich, "Leakage in data mining: Formulation, detection, and avoidance," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2011, pp. 556–563.
[70] J. Engel, J. Gerretzen, E. Szymanska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. C. Buydens, "Breaking with trends in pre-processing?" *TrAC Trends Anal. Chem.*, vol. 50, pp. 96–106, Oct. 2013.
[71] K. Gibert, M. Sànchez–Marrè, and J. Izquierdo, "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining," *AI Commun.*, vol. 29, no. 6, pp. 627–663, Dec. 2016.
[72] M. A. Lones, "How to avoid machine learning pitfalls: A guide for academic researchers," 2021, *arXiv:2108.02497*.
[73] M. Rauschenberger and R. Baeza-Yates, "How to handle health-related small imbalanced data in machine learning?" *I-Com*, vol. 19, no. 3, pp. 215–226, Jan. 2021.
[74] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard, "Navigating the pitfalls of applying machine learning in genomics," *Nature Rev. Genet.*, vol. 23, no. 3, pp. 169–181, Mar. 2022.
[75] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 105524.
[76] D. Singh and B. Singh, "Feature wise normalization: An effective way of normalizing data," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108307.
[77] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[78] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.
[79] A. Shmilovici, *Support Vector Machines*. Boston, MA, USA: Springer, 2010, pp. 231–247.
[80] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.
[81] J. Wu, "Introduction to convolutional neural networks," *Nat. Key Lab Novel Softw. Technol.*, vol. 5, no. 23, p. 495, 2017.
[82] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim, "Mitigating bias in machine learning for medicine," *Commun. Med.*, vol. 1, no. 1, p. 25, Aug. 2021.
[83] I. N. Cofone, "Algorithmic discrimination is an information problem," *Hastings LJ*, vol. 70, p. 1389, Jan. 2018.
[84] T. Hellström, V. Dignum, and S. Bensch, "Bias in machine learning - what is it good for?" 2020, *arXiv:2004.00686*.
[85] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, R. J. Bandara, and K. Michael, "Algorithmic bias in machine learning-based marketing models," *J. Bus. Res.*, vol. 144, pp. 201–216, May 2022.
[86] T. E. Doyle, V. Tucci, C. Zhu, Y. Zhang, B. Yassa, S. Rashidiani, M. A. Khan, R. Samavi, M. Noseworthy, and S. Yule, "Artificial intelligence nomenclature identified from Delphi study on key issues related to trust and barriers to adoption for autonomous systems," 2022, *arXiv:2210.09086*.
[87] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.
[88] Y. Sun, B. Xue, M. Zhang, G. G. Yen, and J. Lv, "Automatically designing CNN architectures using the genetic algorithm for image classification," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3840–3854, Sep. 2020.
[89] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0224365.
[90] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40–79, Jan. 2010.
[91] D. M. Hawkins, S. C. Basak, and D. Mills, "Assessing model fit by cross-validation," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 579–586, Mar. 2003.
[92] F. Maleki, N. Muthukrishnan, K. Ovens, C. Reinhold, and R. Forghani, "Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment," *Neuroimaging Clinics*, vol. 30, no. 4, pp. 433–445, 2020.
[93] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*.
[94] R. Dinga, B. W. Penninx, D. J. Veltman, L. Schmaal, and A. F. Marquand, "Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines," *BioRxiv*, vol. 2019, Aug. 2019, Art. no. 743138.
[95] P. Flach, "Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9808–9814.

RYAN G. L. KOH received the B.Eng. degree in electrical and biomedical engineering from McMaster University, and the Ph.D. degree from the Institute of Biomedical Engineering, University of Toronto, Canada.

He is currently a Postdoctoral Fellow with the KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Canada, and an incoming Postdoctoral Fellow with the Data Science Institute, University of Toronto. His research interests include peripheral nerve interfaces, signal and image processing, and ML for applications in healthcare.

Dr. Koh research has led to numerous awards, such as the NSERC CGS-D Scholarship, the TRI Student Scholarship, and the Sally and Paul Wang Graduate Scholarship in Biomedical Engineering.

MD ASIF KHAN received the B.S. degree in computer science and engineering from North South University, Dhaka, Bangladesh, and the M.A.Sc. degree in electrical and computer engineering from McMaster University, Canada.

From 2021 to 2022, he was a Research Assistant with the Biomedical AI Laboratory. His research involved unsupervised and semi-supervised learning techniques to analyze chronic pain (CP) data, resulting in identification and quantification of co-existing CP mechanisms, which is often unattainable in clinical settings. His research interests include artificial intelligence and machine learning for healthcare applications, especially critical disease diagnosis, prognosis, and treatment. He also possesses a keen interest in the practical implementation and application of data-driven solutions that arise from his research.

SAJJAD RASHIDIANI received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2018, and the M.Sc. degree in electrical and computer engineering from McMaster University, Canada, in 2022.

From 2021 to 2022, he was a Research Assistant with the Biomedical AI Laboratory, McMaster University. He is currently a Data Scientist with Canadian Tire Bank. His research interest includes the development of AI-based decision support systems for hospital readmission.

SAMAH HASSAN received the master's and Ph.D. degrees from the Institute of Medical Science, University of Toronto. She completed her residency in anesthesia and pain medicine with the University of Cairo, Egypt.

During the Ph.D., she developed and validated the new pain competence assessment tool (PCAT), now used at the Inter-faculty Pain Curriculum, University of Toronto. She is currently a Postdoctoral Fellow with The Institute of Education Research (TIER), University Health Network, Canada. She presented her work in numerous national and international forums, including the International Association of Study of Pain (IASP) congress meeting. Her research interests include big data analyses, phenotyping and classification, measurement validity and reliability, and AI integration in healthcare.

VICTORIA TUCCI received the B.H.Sc. degree in health sciences and the M.Sc. degree in global health from McMaster University, Canada. She is currently pursuing the M.D. degree with the University of Toronto.

Her research interests include understanding the trust dynamics between medical AI and healthcare expert end-users. She is exploring the factors that influence trust with these technologies and how they compare to established concepts of trust in the engineering discipline. She is interested in understanding how autonomous systems can be optimized to improve decision-making support and clinician-machine teaming, as well as facilitate the adoption of AI medical technologies into practice. She is passionate about optimizing patient-centered healthcare delivery and understanding how to leverage technology to do so.

THEODORE LIU is currently pursuing the B.H.Sc. degree in health, engineering science and entrepreneurship with McMaster University, Canada.

From 2020 to 2022, he was a Research Assistant with the Biomedical AI Laboratory. His research interests include artificial intelligence in medicine, medical imaging, and literature review. In the future, he hopes to continue to explore and support the adoption of AI medical technologies in clinical practice.

KARLO NESOVIC received the B.Eng. degree in electrical and biomedical engineering from McMaster University, Canada, and the M.A.Sc. degree in biomedical engineering from the Institute of Biomedical Engineering, University of Toronto, Canada, where he is currently pursuing the M.D./Ph.D. degree with the Temerty Faculty of Medicine.

His research interests include peripheral nerve stimulation, autonomic dysfunction, spinal cord injury, sensory processing, and neuromodulation.

DINESH KUMBHARE (Member, IEEE) received the M.Sc. degree in health research management from McMaster University, Canada, and the Ph.D. degree in biomedical engineering from the University of Toronto, Canada.

He is currently a Professor and a Clinician Scientist with the Department of Medicine, University of Toronto, within the Division of Physical Medicine and Rehabilitation and the KITE Research Institute, Toronto Rehabilitation Institute, University Health Network. He is cross-appointed to the Institute of Biomedical Engineering, Faculty of Kinesiology and Physical Education, and the Institute of Health Policy, Management and Evaluation, University of Toronto. He is also an Adjunct Faculty in engineering with McMaster University. He is the Director of the Schroeder Pain Assessment and Rehabilitation Research Centre (SPARC), TRI, a newly established program that will foster a collaborative environment that brings together multidisciplinary and interprofessional constituency of researchers. His research interests include developing more objective measures of pain using ultrasound imaging and quantitative sensory testing.

THOMAS E. DOYLE (Senior Member, IEEE) received the Ph.D. degree from the University of Western Ontario.

From 2014 to 2019, he was the Director of the eHealth Graduate Program, McMaster University, Canada. He is currently an Associate Professor with the Department of Electrical and Computer Engineering and a member of the School of Biomedical Engineering, Faculty of Engineering, McMaster University. In addition, he is a Faculty Affiliate with the Vector Institute for Artificial Intelligence. His research interests include artificial intelligence and ML for human health and performance, human-AI partnership, and trust in autonomous medical advisory systems. His current and past research awards were received from NSERC, CIHR, NSBRI, SOSCIP, MITACS, HHS, and Canadian DND IDEaS.