Sex Parity in Cognitive Fatigue Model Development for Effective Human-Robot Collaboration

Apostolos Kalatzis, Sarah Hopko, Ranjana K. Mehta, Laura Stanley, and Mike P. Wittie

Abstract— **In recent years, robots have become vital to achieving manufacturing competitiveness. Especially in industrial environments, a strong level of interaction is reached when humans and robots form a dynamic system that works together toward achieving a common goal or accomplishing a task. However, the human-robot collaboration can be cognitively demanding, potentially contributing to cognitive fatigue. Therefore, the consideration of cognitive fatigue becomes particularly important to ensure the efficiency and safety in the overall human-robot collaboration. Additionally, sex is an inevitable human factor that needs further investigation for machine learning model development given the perceptual and physiological differences between the sexes in responding to fatigue. As such, this study explored sex differences and labeling strategies in the development of machine learning models for cognitive fatigue detection. Sixteen participants, balanced by sex, recruited to perform a surface finishing task with a UR10 collaborative robot under fatigued and non-fatigued states. Fatigue perception and heart rate activity data collected throughout to create a dataset for cognitive fatigue detection. Equitable machine learning models developed based on perception (survey responses) and condition (fatigue manipulation). The labeling approach had a significant impact on the accuracy and F1-score, where perception-based labels lead to lower accuracy and F1-score for females likely due to sex differences in reporting of fatigue. Additionally, we observed a relationship between heart rate, algorithm type, and labeling approach, where heart rate was the most significant predictor for the two labeling approaches and for all the algorithms utilized. Understanding the implications of label type, algorithm type, and sex on the design of fatigue detection algorithms is essential to designing equitable fatigue-adaptive human-robot collaborations across the sexes.**

ECG, human factors, human-robot collaboration, machine learning, sex differences, robot adaptation

I. Introduction

The emergence of Industry 5.0 places the emphasis on designing robotics that can augment and support human capabilities in manufacturing processes [1], [2]. The combination of the precision and speed of industrial collaborative robots with the creativity and ingenuity of humans can lead to more efficient and safer manufacturing processes, that are otherwise uneconomical or difficult to automate [1] However, the new types of interactions between humans and robots can lead to more complex tasks for operators, shifts in workload, and the emergence of additional human factor considerations, such as trust, situation awareness, and team fluency [3]. Consideration of the resulting impact this human-robot collaboration (HRC) has on the human worker is essential to improve the overall manufacturing systems performance, improve the trustworthiness of HRC designs, and result in a better experience for the operator, where robotics are designed with intelligent support.

A pertinent consideration in such HRC designs is the emergent loading on the operator in the collaborative tasks. Traditionally, a robot's comparative strength provides repetitive support and precision to uniform sub-processes [4]. Allocation of the uniform and nominal events to the robot can result in increased complexity of the operator's tasks. Additionally, the interaction itself with a complex system can similarly increase the load of the operator [5]. Sustaining higher levels of cognitive load can directly lead to faster onsets of cognitive fatigue. Previous research has shown that fatigue is directly associated with increased human error, motivation impairments, and tendency towards complacent and unsafe behaviors [6]–[9]. While the onset of fatigue is undesirable, underloading operators can lead to similar impairments [10]. In some cases, improper design of automation can remove task complexity and underload operators resulting in attention decrements, lower worker satisfaction, and unsafe work conditions [11]. Ensuring appropriate levels of cognitive load is an essential consideration of HRC designs. Regardless of loading levels, the onset of operator's fatigue is an inevitable consideration in HRC designs.

A. Heart Rate Variability as a Fatigue Measure

Providing collaborative robots with the means to perceive their operator's cognitive state surrounding fatigue enables them to make informed decisions to support their operator. This knowledge can be provided to the robot through non-invasive measures sensitive to cognitive fatigue state, such as human physiology. Over the past decades, studies have observed that the autonomic nervous system (ANS) is a physiological indicator of cognitive fatigue [12]–[14]. The ANS consists of two major subsystems: the sympathetic and parasympathetic systems [15]. The sympathetic system is our "fight and flight" response, which activates during the stress, and exertion states. In contrast, the parasympathetic system is our "rest and digest" response, which is dominant during a relaxed state [16]. The most commonly assessed indices of ANS are based on cardiovascular activity [17]. Heart rate variability (HRV) is based on the variations between heartbeats and it has been proven to be a reliable indicator of the ANS's activity [17].

The use of non-invasive covert physiological monitoring

Apostolos Kalatzis, Laura Stanley and Mike P. Wittie are with the Gianforte School of Computing, Montana State University, Bozeman, MT, 59717 USA

Sarah Hopko and Ranjana K. Mehta are with Department of Industrial and System Engineering,Texas AM University, College Station, TX, 77843 USA.

is the first step toward fatigue-adaptive robotics that mitigate the negative effects associated with operator fatigue and allow for fatigue recovery. Machine learning algorithms (ML) can be leveraged to detect cognitive fatigue in real-time. Additionally, research studying fatigue ML can provide novel advancements into understanding physiological differences between operators as well as contextual insights throughout the task.

A frequent consideration for variability between operators is their sex. According to the United States census, 30% of the manufacturing workers have consistently been female [18], yet are frequently not considered in research [19]. Historically, males and females have shown to have different fatigue reporting perceptions [3], [20], as well as different HRV responses to fatigue manipulation [3]. While the consideration of operator sex in fatigue detection ML is essential, operator sex is woefully overlooked in current detection strategies. It is thus unknown how operator sex impacts performance of various fatigue detection model-types, the trade-offs of using fatigue perceptions or condition as model labels, or the physiological predictors that will best detect fatigue in each sex. Moreover, understanding the fundamental sex-differences in perceptual and physiological response can help reveal which types of ML models (e.g., SVM, kNN) will be most successful at explaining the difference between fatigue and no fatigue for each sex as well as inform which labeling strategy provides the best model performance for each sex. It is possible that the underlying sex-differences in perceptual and physiological indicators of fatigue lead to different geometrical properties of HRV data. Thus, this study focuses on sex-parity in the development of fatigue model detections.

B. Machine Learning Detection of Fatigue

While consideration of operator sex is overlooked, previous work has shown the viability of using HRV data with k nearest neighbors (kNN) [21], artificial neural networks (ANN) [22], and random forests (RF) [23] to detect cognitive fatigue. ANNs have been proven to be effective at detecting cognitive fatigue with an accuracy of 91.3% [22]. Additionally, using physiological features extracted from a wrist wearable device has been shown effective at detecting cognitive fatigue [21]. The results of a previous study indicate that a subject independent kNN achieved 75.5% accuracy. Furthermore, based on a recent study RF achieved 57.8% accuracy using three-fold cross-validation and accuracy of 63.9% using principal components analysis and leave-one-out cross-validation [23]. While the previous work indicates the feasibility of detecting fatigue through heart rate features, there is a lack of consideration for the underlying sex-differences in fatigue detection [20], where controlling for sex in sample sizes alone may not be sufficient for equitable ML given the fundamental differences in reporting and responding to fatigue between the sexes, which implicates which model types and labels best classify fatigue.

Perceptual and physiological differences between the sexes have implications for the labeling of machine learning models. There are two main methods of labeling cognitive fatigue: by perception (survey responses) or by manipulation (fatigue/no fatigued condition). The majority of studies exclusively focus on developing machine learning models based on fatigue manipulation [21], [23]. However, it is important to investigate the cost of using perception and condition as ground truth labels given known sex-differences in perceptual and physiological responses so that designs of fatigue detection models are fair between the sexes.

As such, this investigation explores considerations for sex-equitable outcomes in machine learning that detects cognitive fatigue, including understanding the role of sex, labeling approach, and type of ML algorithm. To provide a dataset specific to fatigue in HRC, a human-subjects study was conducted, where manipulation of operator fatigue state as well as a balancing of operator sexes was performed in a popular HRC use case namely, surface finishing. Our primary objectives include:

1) Determine model accuracy and F1 score trade-offs, if any, for labeling models based on perception (survey responses) vs. condition (manipulation) and the interaction with operator sex.
2) Determine model accuracy and F1 score trade-offs, if any, across operator sex (male vs. female) as the main effect.
3) Discuss differences in significant predictors across variables (label type, operator sex, algorithm type).

II. Methods

A. Procedure

Sixteen participants, balanced by sex, were recruited from the university engineering community to perform repetitions of a metal polishing task in collaboration with a UR10 robot (Universal Robots; Denmark). All participants were healthy, right-handed, and had an age distribution of 25.12 ± 3.31 years. This study was approved by the local Institutional Review Board and adhered to COVID-19 safety protocols. Participants were compensated with $70 to attend two sessions, split by the fatigue variable (summing to approximately 7 hours). At each session, as illustrated in Figure 1, participants underwent 40 approximately one-minute trials, in which they used right-handed joystick inputs to navigate the robot's end-effector along a precise S-shaped trajectory. A full repeated measures design was used, where all participants underwent two levels of fatigue, explained below, and two levels of robot assistance [3], to create a dataset used to train cognitive fatigue detection models. This dataset led to a large volume of data per participant and deemed as sufficient to test sex differences given observed differences between men and women in fatigue perceptions with effect sizes greater than 0.2 and sex differences in HRV responses had significant differences with effect sizes greater than 0.3 [3].

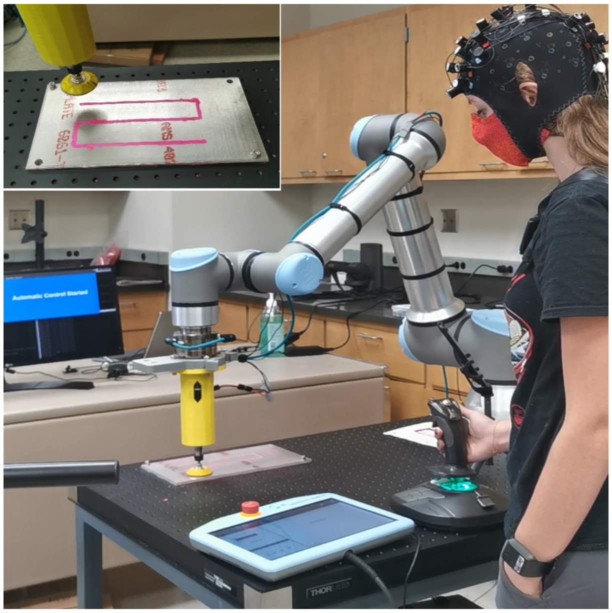

Fig. 1: Surface Polishing Task. Joystick controls are converted to robot joint velocities to navigate an S-shaped trajectory to polish a metal surface.

B. Cognitive Fatigue Manipulation

Participants attended two sessions split by the fatigue variable: no-fatigue vs. fatigue. At the fatigue session, cognitive fatigue was manipulated prior to the participant's interaction with the UR10 robot using a 1-hour computer-based n-back task [3], prior to which participants were familiarized with the task and given a minimum of five minutes to practice. Participants completed a spatial 2-back test which required memorization of the last two locations of stimuli within a 3X3 and pressing the space bar when the current stimulus matched the location of one two-back grid [24]. The spatial n-back test loads spatial working memory and cognition needed to perform a spatial metal polishing task, and the sustained n-back manipulates cognitive fatigue [25], observable in HRV features [26]. Manipulation of fatigue was successfully shown, on average, to increase fatigue perceptions from (2.75 ± 1.73) in the no fatigue trials to (4.73 ±2.47) in the fatigue trials, and to reduce n-back task performance and physiological response. Results from this manipulation are thoroughly discussed elsewhere [26].

C. Dependent Measurements

Two response variables are presented here: fatigue perceptions and heart rate activity measurements. Fatigue perceptions were captured through one question asked after each trial 'What is your level of fatigue?', rated on an interlocked Likert scale (1-very low fatigue, 9-very high fatigue). This measure was selected based on prior n-back work utilizing this metric to monitor fatigue [26]. Heart rate activity measurements were collected using a chest affixed 2-lead electrocardiogram device (Actiheart 5, Camntech, UK). The electrocardiogram (ECG) signal was collected at a sampling rate of 1024Hz. Any ectopic beats or motion artifacts were interpolated [27] and HR and HRV features from the ECG signal for each trial. This study utilized time-, frequency-, and nonlinear-domain metrics of heart rate variability. Previous research indicates that these metrics can be extracted for the trial windows [28] To extract the ECG signal features, the Q wave, R wave, S wave complex was detected for each trial window and then the time series metric of RR interval data was derived.Next, a power spectrum analysis was applied using wavelet transformation, a time-frequency analysis method to scale the decomposed ECG signal into different frequency band signals [29]. Finally, the quality, scaling, and correlation properties of the ECG signal were assessed to extract the nonlinear-domain features. In total, we extracted thirty HRV features in time (mean HR, SD HR, min HR, max HR, mean NN, median NN, RMSSD, SDNN, CVNN, CVSD, NN20, NN50, PNN20, PNN50, range NN, SDSD) [30], frequency (HF, LF, LFnu, HFnu, ratio LF/HF, VLF) [30], and non-linear domain (CSI, modified CSI, CVI, SampEn SD1, SD2, ratio SD2/SD1) [30].

D. Cognitive Fatigue Detection

The next step was to start a machine learning analysis, in which the goal was to detect cognitive fatigue following two different labeling strategies – based on the condition and based on the fatigue perception.

1) Fatigue Based on The Condition: In this case, the machine learning models were trained using the features extracted from the physiological data and the labels were assigned based on the scientific consensus of the 2-back test [24]. A no-fatigue label was assigned at each trial in the no fatigue session and a fatigue label was assigned at each trial on the fatigue session.

2) Fatigue Based on Perception: These models were trained using the features extracted from the physiological data and the labels were assigned based on the self-reported fatigue after each trial [21]. The participants' ratings were thresholded into two classes (fatigue and no-fatigued) on the 9-point rating scales, where the threshold was placed in the middle at 5.

E. Cognitive Fatigue Evaluation

In order to predict the cognitive fatigue, we utilized five machine learning algorithms selected to have linear and non-linear range and models sensitive to smaller sample sizes [31]: a support vector machine (SVM), a kNN (K=5), a logistic regression (LR) classifier, an AdaBoost (AB), and an Random Forest (RF). To determine the best hyperparameters for the classifiers, we applied the grid search method for each combination of the parameters of the models.

To evaluate the classifiers, we trained the models and then performed leave-one-subject-out (LOSO) cross-validation, where data from one participant was randomly selected for testing purposes while data from the other participants were used for training the model. This process was repeated until all the participants were used as the test dataset. We calculated the mean and standard deviation accuracy and the

F1-score for all the participants for each of the labeling approaches we followed. Furthermore, we calculated the overall accuracy and F1-score of the two sexes, where we simply averaged the accuracy and F1-score of the male and female participants to capture the differences when interacting with robots. We also, report the range of accuracy and F1-score of males and females for each ML algorithm.

F. Significant Predictors Calculation

We quantified the HRV predictors of importance using the permutation feature importance structure. This approach is suitable to identify the relationship between the HRV features and the cognitive fatigue outcome by identifying a decrease in the machine learning score every time a single HRV feature randomly shuffles [32]. We applied the permutation feature importance technique for each machine learning model we created under each labeling approach for all participants.

G. Statistical Analysis

Statistical significance was determined through repeated measures analysis of variance (RM ANOVA) tests on accuracy and F1-score. Significance is reported at α = 0.05. The RM ANOVAs were separately run on each model performance metric (Accuracy, F1-score) to test the effects of the three independent variables, label type (condition-based/perception based), machine learning algorithm type (AB, SVM, kNN, LR, RF), and sex (male/female). Separate RM ANOVAs were run on the HRV importance to test the effect of the four independent variables: label type, algorithm type, sex, and HRV feature importance. Post hoc comparisons were performed where needed using Tukey-Test.

III. RESULTS

A. Fatigue Condition Models

1) Accuracy Metric: Table I on the next page represents the mean and standard deviation of accuracy for all participants, males, and females. The accuracy of models evaluated on males and evaluated on females were statistically identical (p=0.915; Figure 2). Using LR the accuracy for males ranged from 7% to 97% and from 32% to 100% for females. Using SVM the accuracy ranged from 22% to 96% for males and from 0% to 100% for females. In the case of RF the accuracy was between 0% and 100% for males and 42% to 95% for females. Using kNN the accuracy ranged from 38% to 78% for males and from 57% to 76% for females. kNN achieved a lower 37% and a higher 99% accuracy for two female participants. Finally, the accuracy of AB was between 0% and 100% for males and between 38% and 81% for females.

2) F1-score Metric: Table I on the next page represents the mean and standard deviation of F1-score for all participants, males, and females. The F1-score of models evaluated on males and evaluated on females were statistically identical (p=0.996; Figure 3). Using LR the F1-score for males ranged from 5% to 99% and from 53% to 100% for females. LR achieved a lower accuracy of 21% for one female participants. Using SVM the F1-score ranged from 40% to 97% for males and from 49% to 98% for females. In the case of RF, the F1-score was between 5% and 100% for males and 31% to 95% for females. Using kNN the accuracy ranged from 53% to 65% for males and from 45% to 84% for females. kNN achieved a lower F1-score of 40% and a higher of 80% for two male participants. Also, the model achieved a higher F1-score of 98% for one female participant. Finally, the F1-score of AB was between 5% and 100% for males and between 7% and 87% for females.

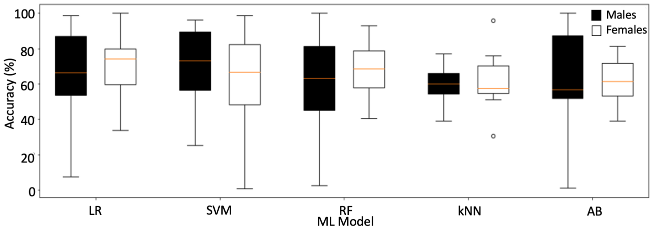

Fig. 2: Accuracy Distribution of Models Evaluated on Males and Females for Condition-Based Models

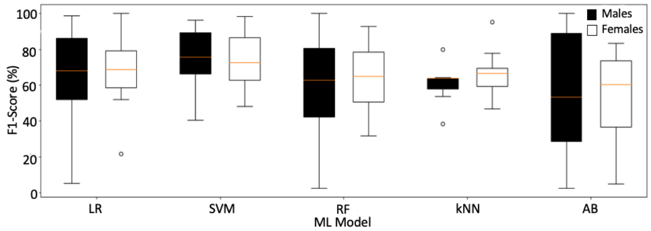

Fig. 3: F1-score Distribution of Models Evaluated on Males and Females for Condition-Based Models

B. Fatigue-Perception Models

1) Accuracy Metric: Table II on the next page represents the mean and standard deviation of accuracy for all participants, males, and females. Models evaluated on males had higher accuracy than evaluated on females (p<0.01; Figure 4). Using LR the accuracy ranged from 33% to 99% for males and from 12% to 70% for females. Using SVM the accuracy ranged from 52% to 100% for males and from 17% to 76% for females. In the case of RF, the accuracy was between 58% and 100% for males and 24% to 63% for females. Using kNN the accuracy ranged from 33% to 100% for males and from 31% and 63% for females. Finally, the accuracy AB was between 38% and 100% for males and between 30% and 64% for females.

2) F1-score Metric: Table II on the next page represents the mean and standard deviation of F1-score for all participants, males, and females for the condition-based machine learning models. Models evaluated on males had higher accuracy than evaluated on females (p<0.01; Figure 5). Using LR the F1-score ranged from 70% to 99% for males. The model achieved a lower F1-score of 24% for one male participant. The F1-score ranged from 21% to 79% for females. Using SVM F1-score ranged from 70% to 100% for males. The model achieved an extremely low F1-score

TABLE I: Mean LOSO CV accuracy (±std) and F1-score (±std) for All participants, Males, and Females for the condition-based Models

Model	ALL		Male		Female	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
AB	61.82±24.10	54.10±31.76	61.79±32.17	56.03±36.03	61.85±14.05	52.17±29.23
SVM	70.93±21.61	74.28±17.36	69.20±24.4	74.54±18.32	71.81±21.12	73.30±17.66
kNN (k=5)	60.42±15.71	63.67±13.48	59.20±11.98	60.90±11.68	61.65±19.12	66.36±15.35
LR	66.46±25.80	65.27±26.13	64.28±30.63	64.83±30.39	68.64±21.86	66.24±23.20
RF	64.74±24.77	64.23±25.66	62.85±31.23	64.63±30.57	69.28±15.26	67.41±18.93

AB = AdaBoost, SVM = Support Vector Machine, RBF = Radial Basis Function, kNN = k Nearest Neighbor, LR = Logistic Regression, RF = Random Forest

TABLE II: Mean LOSO CV Accuracy (±std) and F1-score (±std) for All participants, Males, and Females for the Fatigue perception-based Models

Model	ALL		Male		Female	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
AB	64.94±22.29	64.21±32.14	74.66±24.15	77.37±29.85	50.31±11.76	51.68±12.69
SVM	64.69±25.76	66.65±31.76	79.22±21.33	78.30±21.80	50.16±22.02	55.00±26.57
kNN (k=5)	64.94±25.99	69.38±27.31	81.08±23.04	83.93±25.36	51.80±11.53	50.19±13.31
LR	61.58±27.08	69.92±25.94	76.55±23.48	81.57±24.8	46.61±22.53	58.28±23.30
RF	65.60±21.61	65.15±21.94	79.40±20.79	78.05±21.67	51.13±12.2	51.12±13.44

AB = AdaBoost, SVM = Support Vector Machine, RBF = Radial Basis Function, kNN = k Nearest Neighbor, LR = Logistic Regression, RF = Random Forest

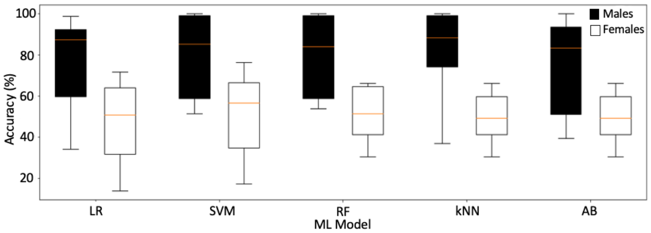

Fig. 4: F1-score Distribution of Models Evaluated on Males and Females for Condition-Based Models

of 0% for one male participant. The F1-score was from 7% to 78% for females. In the case of RF, the F1-score was between 50% and 100% for males and 35% to 70% for females. Using kNN the accuracy ranged from 75% to 100% for males. The model achieved a lower F1-score of 23% for one male participant. The F1-score ranged from 36% to 70% for females. Finally, the F1-score of AB was between 64% and 100% for males. The model achieved an extreme low F1-score of 8% for one male participant. The F1-score was between 31% and 78% for females.

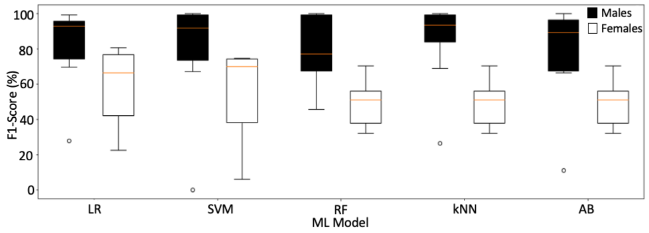

Fig. 5: F1-score Distribution of Models Evaluated on Males and Females for Condition-Based Models

C. Comparison and Differences Across Sexes

1) Accuracy Metric: A statistically significant main effect of sex on accuracy was observed ($p<0.01$, η^2=0.08) with greater accuracy for models evaluated on male participants (70.32% ± 24.14%) than female (58.32% ± 17.28%), regardless of label. There was also a significant interaction effect of sex and label type ($p<0.01$, η^2=0.08)). The post-hoc analysis reviewed; in perception-based models, models evaluated on males had higher accuracy (78.18% ± 22.75%) than females (50% ± 16.28; $p<0.01$). Furthermore, models evaluated on males achieved higher accuracy using perception-based labels compared to condition-based labels (p=0.01). Condition-based labels led to higher accuracy for females (66.65% ± 18.28) compared to perception-based labels (50% ± 16.28; $p<0.01$). Finally, male condition-based models achieved higher accuracy (62.46% ± 26.08) compared to female perception-based (50% ± 16.28; p=0.04). However, there were no significant differences on comparing female condition-based to male condition-based and female condition-based to male perception-based (all $p>0.08$). All other main effects and interactions were not significant (all $p>0.830$).

2) F1-score Metric: A statistically significant main effect of sex on F1-score was observed ($p<0.01$, η^2=0.066)) with greater F1-score of models evaluated on male participants (72.11% ± 25.18%) than female (59.17% ± 18.07%), regardless the label. There was also a significant effect of sex and label type ($p<0.01$,η^2=0.079). The post hoc analysis reviewed; in perception-based models, models evaluated on males had higher F1-score (79.84% ± 24.96% than females (53.25% ± 17.86%; $p<0.01$). Furthermore, models evaluated on males achieved higher F1-scores (79.84% ± 24.96% using perception-based labels compared to condition-based labels (64.38% ± 25.40%; p=0.01). Finally, male perception-based models achieved higher F1-

scores (79.84% ± 24.96% compared to female condition-based (65.09% ± 18.28%; p=0.03). However, there were no significant differences in comparing female condition-based to male condition-based, female perception-based to male condition-based and female perception-based to female condition-based (all p>0.22). All other main effects and interactions were not significant (all p>0.860).

D. Significant Predictors

A statistically significant main effect of sex (p=0.01), model type (p<0.01), label type (p<0.01), and HRV feature (p<0.01) was observed. There was also a significant effect of model type and label type (p<0.01, η^2=0.02). From the Tukey post-hoc analysis we observed that irrespective of the model type, condition-based models attained higher importance means which is likely driven by having less noisy HRV features.

There was also a significant effect of model type and HRV features (p<0.01, η^2=0.49). For the two ensemble models (AB, RF) Mean HR was the most significant predictor of cognitive fatigue. Mean HR was significantly different from the rest of the models and HRV features (p<0.01). For LR and kNN, pNN20 was the significant predictor. However, for SVM, the most important predictor was median NN. There was also a significant effect of label type and HRV features (p<0.01, η^2=0.38). For the condition-based models median NN and Mean HR were the most significant predictors. For the perception-based models, the most important feature was the Mean HR indicating that irrespective of label type mean HR appeared to contribute the most at detecting cognitive fatigue. Another significant predictor that was common for both labeling approaches was the minimum HR. Although, there was no interaction effect of sex and HRV features (p=0.75, η^2=0.03). We performed a post hoc to identify if there were any common significant predictors. Interestingly, we observed that irrespective of the sex, mean HR was the most significant predictor followed by median NN, then pNN20 and minimum HR.

IV. Discussion

This study investigated sex-equitable cognitive fatigue detection. This was done through the utilization of HRV features from an HRC experiment in which participants performed a collaborative task under fatigued and non-fatigued conditions. Five machine learning algorithms were utilized for detecting cognitive fatigue. For each algorithm, we explored two different labeling approaches; one based on fatigue perception and another on fatigue condition. Then tested the effects of labeling approach, model type, and sex on the accuracy and F1-score of the models. The key takeaways of the study include:

1) Labeling ML models by fatigue perception vs. fatigue manipulation yields different accuracy and F1-scores.
2) Perception-based models trained on data collected from females have reduced accuracy and F1-score.
3) Mean HR is a significant predictor of cognitive fatigue.

A. Labeling ML Models by Fatigue Perception vs. Fatigue Manipulation Yields Different Accuracy and F1-score

Depending on the labeling approach, we notice differences in the overall accuracy and F1-score of the machine learning algorithms for detecting cognitive fatigue. The two ensemble machine learning models (i.e., AB, RF) and kNN achieved higher mean accuracy and F1-score using perception-based labels. On the other hand, SVM and LR performed better using the condition-based labels for all the participants. These results indicate that condition-based labels created linear separable data which can explain the higher performance scores of the linear classifiers (i.e., LR, SVM). Previous research indicates that there is a direct relationship between cognitive fatigue and HRV [12]. The predominant activity of the autonomic nervous system during the fatigue state turned to the sympathetic activity from parasympathetic activity [14]. This caused differences in the HRV features that could construct a decision boundary hyperplane that divides the two classes (fatigue, no fatigue) when using condition-based labels. On the other hand, the higher performance of the non-linear algorithms using perception-based labels indicates that perceptions do not follow a linearly separable cluster. These results indicate that the decision to label a model based on condition or perception should follow the choice of the appropriate ML algorithm. Nonlinear ML algorithms are a better fit when a dataset is labeled using subjective perception labels whereas linear models should be utilized in the case of the condition-based labeling approach.

B. Models Trained on Data Collected From Female Participants Have Reduced Accuracy and F1-scores For Perception Models

Using perception-based ground truth labels, models achieved significantly lower accuracy and F1-score evaluated on female participants compared to condition-based ground truth labels. Accuracy was significantly lower across all models with SVM showcasing the highest difference and AB showcasing the lowest difference. For F1-score we observe the same pattern, all models achieved lower scores using perception-based labels. Again, SVM achieved the highest F1-score differences while AB had the smallest gap with F1-score to be close using either label. The low accuracy and F1-score of the perception-based models can be attributed to the way males and females report cognitive fatigue. Historically, the subjective responses of cognitive fatigue vary by sex [33], where females tend to report increased levels of fatigue than males. In this study, all the algorithms were able to capture the underlying patterns of cognitive fatigue in the case of the male participants [3]. It is likely that for this study male participants reported cognitive fatigue more accurately with respect to their physiological data. On the other, the perceived fatigue perception had greater variance for female participant than male. This resulted in heterogeneous datasets with unclear class boundaries. These findings highlight that condition-based labels are more appropriate to accommodate and address sex-differences in HRC. Models evaluated on males achieved the highest accuracy and F1-score using perception-based labels. However, the poor accuracy and F1-score of models evaluated on females indicate that condition-based labels lead to more equitable machine learning algorithms.

C. HR is a Significant Predictor of Cognitive Fatigue

The statistical analysis showed that mean heart rate (HR) contributes the most to predicting cognitive fatigue. Irrespective of the labeling approach and sex, HR seems to be the most significant predictor. This is in line with previous literature that has reported a positive correlation between HR levels and cognitive fatigue [34]. This result indicates using HR as a measure of cognitive fatigue can lead to the development of more generalizable machine learning models of cognitive fatigue in HRC, serving as a moderator to account for varying levels of cognitive fatigue across different labeling approaches and operator sex. Additionally, the recent development of new wearable sensor technologies, such as smartwatches, has offered unobtrusive data collection that can provide a highly deployable objective measures for cognitive fatigue detection [33]. Cognitive fatigue can be monitored during HRC tasks by using wearable devices that do not limit the freedom of movement and, hence, could be truly used in real operative scenarios where a human is interacting with a robot.

D. Sex-Balanced Sampling is Not Sufficient for Designing Equitable Models

In this study, we carried out sex-balanced sampling for the detection of cognitive fatigue. The results indicate that the development of equitable machine learning models requires careful consideration of the labeling approach and the algorithm type. The overall accuracy and F1-score was not significantly different for the two labeling approaches. However, using perception-based models we observed significant differences in accuracy and F1-score between the sexes [3], [33]. Specifically, models evaluated on female participants had drastically reduced accuracy and F1-score irrespective of the algorithm type used. This has implications on designing algorithms that are built on perception due to differences in how males and females report fatigue [3], [20]. Studies have shown that females tend to be more concerned about task achievement than males and often have a lower expectation of success [35], [36]. These studies suggest that female participants perceived greater fatigue magnitude than male participants would and therefore incorrectly aligned their physiological responses to higher fatigue perceptions [35], [36]. Additionally, we observed that the choice of the machine learning algorithms seems to lead to different accuracy and F1-score for males and females. We observed cases where one algorithm performs better at detecting cognitive fatigue for males while the same algorithm leads to poor performance results for females. These results indicate that a more comprehensive exploration might be needed for the development of equitable machine learning models where separate algorithms should utilize for the sexes.

E. Machine Learning Supports Workers Under Cognitive Fatigue

Adopting the observations of this study can lead to the design more effective HRCs. By detecting the cognitive fatigue state of a worker through machine learning support, we can improve their ability to perform tasks more efficiently and safely. A promising way is tuning the interaction with the robot depending on the operator's cognitive fatigue state by implementing robot control architecture that allows the robot to detect and respond to the cognitive fatigue needs of its collaborator. However, the imperfect accuracy, and consequently the presence of false cognitive fatigue alarms should be considered in order to achieve a calibrated trust and reliance on automation [37]. Trust and reliance should be part of a closed-loop which involves various factors, including individual, organizational, cultural, and environmental context that impact the trust evolution and reliance on automation [38].

F. Study Limitations and Future Work

One of the limitations of this study is the use of a singular question to report cognitive fatigue which can result in partial understanding and reporting of fatigue perceptions. There is a trade-off of utilizing larger composite surveys as these are more invasive to the task thus allowing for less frequent monitoring; however, future work should consider the accuracy of responses to surveys as the ground truth for someone's fatigue. This limitation in itself provides the foundation for discussion on the use of any subjective response as operator's fatigue state. Another limitation is that the participants recruited in this study were college students predominately seeking advanced degrees in engineering. Future work should focus on industry workers as the majority of jobs in manufacturing are taken up by high-school graduates [39]. Furthermore, more experienced workers may have different strategies to compensate for fatigue, such as talking to coworkers or offloading work [5], and may report fatigue levels differently. Efforts should be made to ensure a representative sample of younger and older, and male and female workers in the planned participant pool. Similar to the need to consider sex-differences in model development, age of the operator will impact how fatigue impacts perceptual and physiological indicators, thus future work should consider age-parity in model development. Furthermore, while a large volume of data was collected for each sample, a sample size of eight males and eight females is susceptible to recruiting outliers, thus future work should utilize larger between-subjects sample sizes. Finally, this study focused on the detection of cognitive fatigue in an offline setting using a dataset that was constructed from an HRI case study. The findings presented in this study will be adopted for the next stage of our research where we will feed cognitive-state HRV data to machine learning algorithms to detect worker cognitive fatigue state. Our approach will include a real-time loop where the robot adapts its behavior (e.g., stop, slow, or prompt for action and remote supervision). based on the operator's cognitive fatigue state.

V. Conclusion

This study investigated the development of equitable machine learning models for cognitive fatigue detection. The developed machine learning models were evaluated based on the LOSO cross-validation and the performance of the models investigated under two different labeling approaches (i.e., induced fatigue, perception). Our findings indicate that perception-based labels led to lower accuracy and F1-Score for models trained on data collected on female participants. Models trained on data collected from male participants achieved higher accuracy and F1-Score using perception-based ground truth labels. Condition-based labels led to the development of more sex equitable machine learning models where accuracy and F1-Score are statistically identical and HR is equally important for the development of the models. These findings demonstrate that practitioners should consider sex differences to develop more equitable ML models in order to achieve more effective HRCs.

References

[1] A. Khalid, P. Kirisci, Z. Ghrairi, K. Thoben, and J. Pannek, "Towards implementing safety and security concepts for human-robot collaboration in the context of Industry 4.0," in *MATADOR Advanced Manufacturing*, Jul. 2017.

[2] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0 – Inception, conception and perception," *Manufacturing Systems*, vol. 61, pp. 530–535, Oct. 2021.

[3] S. K. Hopko, R. Khurana, R. K. Mehta, and P. R. Pagilla, "Effect of Cognitive Fatigue, Operator Sex, and Robot Assistance on Task Performance Metrics, Workload, and Situation Awareness in Human-Robot Collaboration," *Robotics and Automation Letters*, vol. 6, no. 2, pp. 3049–3056, Apr. 2021.

[4] A. Vysocky and P. Novak, "Human-robot collaboration in industry," *MM Science Journal*, vol. 9, no. 2, pp. 903–906, Jun. 2016.

[5] L. Lu, F. M. Megahed, R. F. Sesek, and L. A. Cavuoto, "A survey of the prevalence of fatigue, its precursors and individual coping mechanisms among US manufacturing workers," *Applied ergonomics*, vol. 65, pp. 139–151, Nov. 2017.

[6] S. Sonnentag and F. R. Zijlstra, "Job characteristics and off-job activities as predictors of need for recovery, well-being, and fatigue," *Journal of applied psychology*, vol. 91, no. 2, p. 330, Mar. 2006.

[7] N. W. Van Yperen and O. Janssen, "Fatigued and dissatisfied or fatigued but satisfied? Goal orientations and responses to high job demands," *Academy of Management Journal*, vol. 45, no. 6, pp. 1161–1171, Nov. 2002.

[8] J. Dorrian, S. D. Baulk, and D. Dawson, "Work hours, workload, sleep and fatigue in Australian Rail Industry employees," *Applied ergonomics*, vol. 42, no. 2, pp. 202–209, Jan. 2011.

[9] M. R. Grech, A. Neal, G. Yeo, M. Humphreys, and S. Smith, "An examination of the relationship between workload and fatigue within and across consecutive days of work: Is the relationship static or dynamic?" *Journal of occupational health psychology*, vol. 14, no. 3, p. 231, Jul. 2009.

[10] P. Broadhurst, "The interaction of task difficulty and motivation: The Yerkes Dodson law revived," *Acta Psychologica*, vol. 16, pp. 321–338, 1959.

[11] A. Meissner, A. Trübswetter, A. S. Conti-Kufner, and J. Schmidtler, "Friend or foe? understanding assembly workers' acceptance of human-robot collaboration," *Transactions on Human-Robot Interaction*, vol. 10, no. 1, pp. 1–30, Jul. 2020.

[12] N. Egelund, "Spectral analysis of heart rate variability as an indicator of driver fatigue," *Ergonomics*, vol. 25, no. 7, pp. 663–672, Jul. 1982.

[13] F. Ding, N. Fu, S. Alsamarai, and Y. Xu, "Correlating heart rate variability with mental fatigue," *Worchester Polytechnic Institute*, Apr. 2012.

[14] H. M. Melo, L. M. Nascimento, and E. Takase, "Mental fatigue and heart rate variability (HRV): The time-on-task effect," *Psychology & Neuroscience*, vol. 10, no. 4, p. 428, Dec. 2017.

[15] A. D. Loewy and K. M. Spyer, *Central regulation of autonomic functions*. Oxford University Press, 1990.

[16] R. A. Rhoades and D. R. Bell, *Medical phisiology: Principles for clinical medicine*. Lippincott Williams & Wilkins, 2012.

[17] P. Brodal, *The central nervous system: structure and function*. oxford university Press, 2004.

[18] US Census Bureau, "Women in manufacturing," https://www.census.gov/newsroom/blogs/random-samplings/2017/10/women-manufacturing.html, Oct. 2020.

[19] S. Hopko, J. Wang, and R. Mehta, "Human factors considerations and metrics in shared space human-robot collaboration: A systematic review," *Frontiers in Robotics and AI*, vol. 9, 2022.

[20] P. J. Caplan, M. Crawford, J. S. Hyde, and J. T. Richardson, *Gender Differences in Human Cognition. Counterpoints: Cognition, Memory, and Language Series*. ERIC, 1997.

[21] S. Huang, J. Li, P. Zhang, and W. Zhang, "Detection of mental fatigue state with wearable ecg devices," *International journal of medical informatics*, vol. 119, pp. 39–46, Oct. 2018.

[22] H. Al-Libawy, A. Al-Ataby, W. Al-Nuaimy, and M. A. Al-Taee, "HRV-based operator fatigue analysis and classification using wearable sensors," in *Multi-Conference on Systems, Signals & Devices*, Mar. 2016.

[23] K. Tsunoda, A. Chiba, K. Yoshida, T. Watanabe, and O. Mizuno, "Predicting changes in cognitive performance using heart rate variability," *Transactions on Information and Systems*, vol. 100, no. 10, pp. 2411–2419, Oct. 2017.

[24] J. F. Hopstaken, D. Van Der Linden, A. B. Bakker, and M. A. Kompier, "A multifaceted investigation of the link between mental fatigue and task disengagement," *Psychophysiology*, vol. 52, no. 3, pp. 305–315, Mar. 2015.

[25] X. Caseras, D. Mataix-Cols, V. Giampietro, K. A. Rimes, M. Brammer, F. Zelaya, T. Chalder, and E. L. Godfrey, "Probing the working memory system in chronic fatigue syndrome: a functional magnetic resonance imaging study using the n-back task," *Psychosomatic Medicine*, vol. 68, no. 6, pp. 947–955, Nov. 2006.

[26] R. Karthikeyan, J. Carrizales, C. Johnson, and R. Mehta, "Visuospatial working memory under fatigue: Observations with cerebral hemodynamics and heart rate variability," *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2021.

[27] V. Marked, *Correction of the heart rate variability signal for ectopics and missing beats*. Futura Publishing Company, 1995.

[28] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *Engineering in medicine and biology society*, Feb. 2007.

[29] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Applied ergonomics*, vol. 74, pp. 221–232, Jan. 2019.

[30] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[31] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *International conference on Machine learning*, Jun. 2006.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[33] W.-L. Huang, L.-R. Chang, T. B. Kuo, Y.-H. Lin, Y.-Z. Chen, and C. C. Yang, "Gender differences in personality and heart-rate variability," *Psychiatry research*, vol. 209, no. 3, pp. 652–657, Oct. 2013.

[34] C. Zhang and X. Yu, "Estimating mental fatigue based on electroencephalogram and heart rate variability," *Polish Journal of Medical Physics And Engineering*, vol. 16, no. 2, p. 67, Jan. 2010.

[35] N. E. Betz and L. F. Fitzgerald, *The career psychology of women.* Academic Press, 1987.

[36] S. Pyke and S. Kahili, "Sex differences in characteristics presumed relevant to professional productivity," *Psychology of women quarterly*, vol. 8, no. 2, pp. 189–192, Dec. 1983.

[37] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, Jun. 2003.

[38] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, Mar. 2004.

[39] US Census Bureau, "Occupational employment projections to 2022," https://www.bls.gov/opub/mlr/2013/article/pdf/occupational-employment-projections-to-2022.pdf, Dec. 2013.