Human Factors in an Interactive Machine Learning System

Cedric Bone
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, U.S.A.
cb9017@rit.edu

***Abstract*—Machine learning relies on data. Interactive machine learning (IML) is a promising paradigm that combines machine learning algorithms with human intelligence to create a more efficient learning process. In these systems, considering both the algorithm and the human factors involved is essential to creating an effective system. While there has been extensive research on the algorithmic factors of IML systems, there has been less effort in studying human factors in IML system development. In particular, cognitive load is essential when designing systems humans interact with because high cognitive load impacts the user experience and can lead to decreased performance. The focus of this study is to determine the effect of *collaboration* and *system controllability* on participant cognitive load and model performance in an IML labeling system. This study uses an IML system that links a long short-term memory (LSTM) model to a named entity recognition (NER) labeling interface. The study is still ongoing but preliminary analysis indicates that pair-based collaboration may not impact cognitive load or model performance. Conversely, the controllability of the system may impact cognitive load, but not model performance.**

***Index Terms*—Human Factors, Interactive Machine Learning, Cognitive Load, Human-Computer Interaction**

I. Introduction

This research focuses on interactive machine learning (IML). IML is a subfield of machine learning (ML) that intersects with human-computer interaction (HCI). HCI is a multidisciplinary field that studies the design, evaluation, and implementation of interactive computing systems for human use to improve user experience and the effectiveness of computing systems [1].

In IML systems, the collaboration between algorithms and humans introduces a dynamic aspect to the learning process. This study aims to provide insights that will inform the design and development of future IML systems. Specifically, it aims to understand the impact of *pair-based collaboration* and varying *system controllability* on cognitive load and system performance in an IML system. Cognitive load is important to consider in IML systems because it can impact how effectively humans interact with these systems. Cognitive load theory asserts that when dealing with new information, people have a limited capacity for processing information, which, when exceeded, can lead to decreased performance [2]–[4]. In the context of IML, where users perform decision-making tasks, managing cognitive load is important to prevent overload. High cognitive load can impair a user's ability to perform tasks, reducing the effectiveness of the human-computer interaction and potentially negatively affect the learning of the ML model. By understanding cognitive load, designers of IML systems can make more user-friendly interfaces that enhance user performance.

In this study, the NASA-TLX rating scale survey, pupillometry, and galvanic skin response (GSR) are used to collect subjective and objective indications of cognitive load, while the F1 score indicates model performance. By including cognitive load measurements such as the NASA-TLX rating scale survey, pupillometry, and GSR, this study assesses users' perceived workload when interacting with the IML system, which impacts their ability to make accurate annotations. The F1 score is chosen as a performance metric for the model because it combines the precision and recall of the model's predictions giving a metric where both false positives and false negatives carry importance.

This study specifically examines pair-based collaboration and system controllability to find their roles in managing cognitive load and enhancing system efficacy. Pair-based collaboration could distribute cognitive tasks more evenly, potentially reducing individual cognitive burden and mitigating reduced annotator performance [2]. System controllability allows users to influence model behavior directly, which could increase cognitive load due to the added complexity of control mechanisms.

This paper seeks to answer:

- **RQ 1:** How does pair-based collaboration affect cognitive load and model performance in an IML system?
- **RQ 2:** What is the effect of varying the controllability of the system on cognitive load and model performance in an IML system?

To conduct the study, an IML system was developed linking a named-entity recognition (NER) labeling interface to a long short-term memory (LSTM) neural network. NER was chosen as the IML annotation task due to its intuitive nature and ability to easily adjust the difficulty by altering the text and set of entity categories. We chose an LSTM model for the machine learning algorithm because of its proficiency in processing and learning sequences of data like text, making it suited for tasks like NER. This IML system serves as the experimental ground for the study. In the subsequent sections, this paper provides background on the existing literature, emphasizing the need to consider human factors in IML systems, and outlines the methods used in the study, with preliminary explorative results.

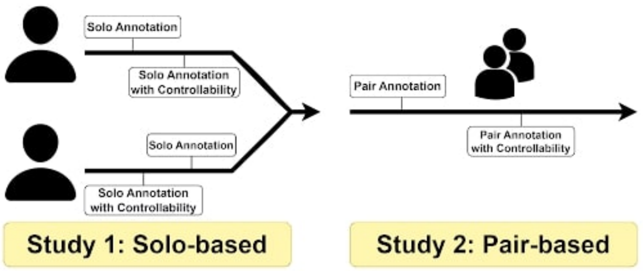

Solo Annotation
Solo Annotation with Controllability
Solo Annotation
Solo Annotation with Controllability
Pair Annotation
Pair Annotation with Controllability
Study 1: Solo-based
Study 2: Pair-based

Fig. 1: Diagram of the experimental setup for solo and pair studies, showing the counterbalanced design where participants either begin with or without system controllability. Participants initially engage in the solo study, followed by the option to return for the paired study, which is also counterbalanced to have the starting controllability condition alternate.

II. Critical Literature Review

Human knowledge and algorithmic efficiency are essential components of IML systems. This literature review covers some essential aspects of IML, focusing on human factors, specifically cognitive load, and the effects of collaboration on cognitive load. Both are important for understanding the potential uses and challenges of IML systems.

A. *Interactive Machine Learning*

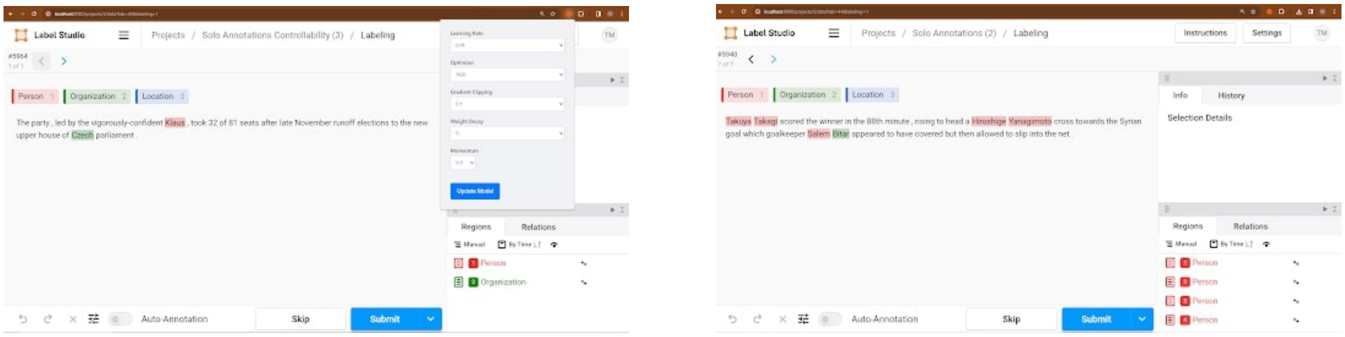
IML is a framework in which interfaces are designed so that users can perform actions that allow an ML algorithm to learn from their guidance [5]. IML systems are useful in scenarios with limited annotated data or where algorithmic predictions could benefit from human insights, such as in medical diagnosis, cybersecurity, and affect modeling. In medical diagnosis, IMLs allow medical professionals to refine disease detection algorithms by providing expert feedback on prediction outcomes, allowing for non-ML experts to train domain-specific models, increasing the speed and reliability of annotations [6]. Similarly, in cybersecurity, analysts use IML systems to adaptively train models on evolving threat patterns, improving the models' ability to detect new attacks [7], [8]. The adaptability of IML systems also makes them suited to tasks that require a contextual understanding. This ability is useful for tasks like affect modeling, where a model can be tuned based on changes in user's biophysical and behavioral responses [9] and they can offer a personalized experience that improves over time with user interaction [10].

Since IML systems combine the capabilities of machine learning algorithms and the intelligence of human collaborators with human-agent teaming, they are suited to address some of the shortcomings of fully autonomous ML systems. One of the main shortcomings of many modern ML algorithms is the requirement for large amounts of annotated data [11]–[13]. The cost associated with data is a challenge, necessitating the exploration of more efficient learning processes. Semi-supervised learning addresses this issue by using both labeled and unlabeled data in training. However, it faces challenges, such as the risk of reinforcing incorrect assumptions if the unlabeled data does not accurately represent the underlying distribution, or if the data has noisy labels [14], [15]. Unsupervised learning does not rely on labeled data. It tries to find latent structures in data. This approach also has challenges, including the absence of target values for validation, which can complicate evaluation. It also often results in models that are difficult to interpret [16]–[19]. Like semi-supervised algorithms, IML systems can also reduce dependency on annotated datasets [6], [20]. IML systems address some of the issues with semi-supervised and unsupervised learning such as mitigating the risk of reinforcing incorrect assumptions. By incorporating human insights into the learning process, IML helps guide the model to be more reliable, improving cases where data is scarce, noisy, or non-representative. Also, IML improves model transparency by allowing users to influence the learning process directly. This involvement not only helps in fine-tuning the algorithms but also enhances user trust in the system [21].

In IML, users and algorithms can solve challenging problems together [22]. This collaboration in IML systems is important because it can lead to better performance than a system without a human-in-the-loop, along with allowing personalization [23], [24]. However, human factors in IML systems have not had the same depth of exploration as the algorithmic factors. Many of the current publications simulate human participants instead of actually having humans in the loop, and the publications that do include actual humans usually indirectly measure human factors through model perforce or subjective measures [7], [25]. This study takes steps toward addressing this gap.

B. *Human Factors*

There are a wide range of human factors in IML systems that can significantly influence the interaction between humans and machine learning technology. Understanding these factors is needed for designing IML systems that are not only efficient but also user-friendly and accessible to a diverse range of users [26]. There are many human factors to consider when designing an IML system including fatigue, motivation, expertise, expectations, and agency [27]–[29]. When a user is subjected to a task needing prolonged attention and decision-making, fatigue becomes a concern. Fatigue can significantly affect user performance in data annotation or problem-solving tasks within IML environments [30], [31]. Fatigue in the context of IML systems refers to the mental or physical exhaustion that users experience after prolonged engagement with annotation or classification tasks. There are many effects that fatigue can have on humans in IML systems. Fatigue can lead to decreased response times and reduced performance. Psychologically, it can diminish user motivation and engagement, potentially leading to a higher dropout rate among participants in long-term studies or projects [32]. To mitigate the effects of fatigue,



The party , led by the vigorously-confident Klaus , took 32 of 81 seats after late November runoff elections to the new upper house of Czech parliament .

Takuya Takagi scored the winner in the 88th minute , rising to head a Hiroshige Yanagimoto cross towards the Syrian goal which goalkeeper Salem Bitar appeared to have covered but then allowed to slip into the net .

(a) The IML's NER labeling interface with the control panel active, allowing users to update parameters that control how the LSTM model gets tuned on user annotations.

(b) The IML's NER labeling interface without the control panel active. Users do not control how the model is tuned from their anotations

Fig. 2: Comparison of the IML system's NER labeling interface showing system controllability.

some strategies may be implementing adaptive interfaces that adjust task difficulty based on real-time assessment of user performance and fatigue levels that introduce mechanisms for regular breaks or variations in task type to sustain engagement and performance over time [4]. Motivation drives users to engage in IML tasks investing their time and cognitive resources into the tasks. Intrinsic motivation can be encouraged by providing meaningful feedback, a sense of progress, and opportunities for autonomy [33]. Extrinsic motivators, such as rewards or recognition, can also play a role in keeping a user participating and giving effort [34]. When a user is motivated it can lead to better annotation quality [35]. Expertise and expectations affect how users interact with IML systems, with experts and beginners having different preferences and expectations. Expert users may require more control and flexibility, while less experienced users may benefit from guidance and simplified interfaces. Understanding the user's level of expertise is important for designing interactive systems that cater to varying needs [26], [28], [36], [37].

IML systems require users to complete problem-solving tasks, usually involving annotation or classification of data. These tasks can impose mental effort on working memory [38]. Cognitive load theory creates a framework for understanding how people process information during problem-solving [39]. High cognitive loads can impair user performance, reduce the effectiveness of learning, and lead to errors [3], [4]. Cognitive overload occurs when the demands of a task exceed a user's cognitive capacity, leading to decreased task performance and increased frustration [40]. The concept of cognitive load is particularly relevant to HCI because it influences how users interact with technology. In IML systems, designers must balance the complexity of the tasks with the users' ability to process information effectively.

Cognitive load can be estimated by subjective measures or indicated by biophysical responses. The NASA-TLX survey is a widely used tool that provides a subjective measure of perceived workload. It assesses perceived workload across six dimensions (*mental demand*, *physical demand*, *temporal demand*, *performance*, *effort*, and *frustration*) [41], [42]. In addition, various forms of biophysical responses can be used to indicate cognitive load such as heart rate variability, blink rate, pupil dilation, and skin conductance. Heart rate variability, for example, can reflect changes in cognitive load and stress levels [43], [44]. Blink rate and pupil dilation are also indicative of cognitive load, with increased blink rate and pupil size being associated with higher levels of mental effort [45]–[47]. Skin conductance, or GSR, measures the electrical conductance of the skin, which varies with sweat production and can indicate emotional arousal and cognitive load [48], [49]. By integrating both subjective and objective measures of cognitive load, a deeper understanding of how IML systems impact users can be formed.

The interaction between collaboration and cognitive load in IML systems is a complex dynamic that can significantly affect user performance. Collaboration, when effectively integrated into IML systems, could possibly distribute cognitive load among users [2], and may strengthen annotation robustness. However, the effectiveness of collaboration is dependent on the design of the system and the nature of the task. That means a nuanced understanding of how collaborative efforts influence cognitive load and system performance is needed. Collaborative learning theories suggest that working in groups allows for the distribution of cognitive tasks among members, potentially reducing the cognitive load on individual users. From a cognitive load theory (CLT) perspective, this distribution can transform an otherwise high cognitive load task into a series of lower load tasks, making the information processing more manageable [50]. However, collaboration introduces its own set of cognitive demands like those imposed by transactive activities, including coordination costs and the need for effective communication, which can, in some cases, offset the benefits of collaboration [51]. Collaboration may even enhance motivation, which is important for sustaining participation in IML tasks [52]. However, the impact of collaboration on cognitive load is mixed, with some studies reporting reduced perceived effort and improved performance, while others showing potential increases in cognitive load due to the complexities of group dynamics and task coordination

[2]. This study contributes to expanding the insights about the relationship between cognitive load and collaboration in hopes of improving our understanding of designing IML interfaces.

III. Research Plan

A. Research Agenda

Since the objective of this study is to explore human factors in IML systems, particularly cognitive load, a systematic investigation of the effects of pair-based collaboration and system controllability is needed. By using a factorial experimental design, the human-computer interaction within this IML environment can be better understood. It is hypothesized that pair-based collaboration will mitigate cognitive load and enhance model performance by distributing the induced cognitive load. Conversely, it is expected that increased system controllability will raise cognitive load due to the added complexity and decision-making it introduces, possibly at the expense of model performance. To test the hypotheses an IML system was implemented.

B. Methodology

The IML system was implemented with the Label Studio open-source data labeling tool as the front-end (Figure 2) which was connected to a backend controlled by an LSTM model. LSTMs are a type of recurrent neural network (RNN) capable of learning long-term dependencies [53]. That is important in NER tasks where the context and sequence of words influence their classification. An LSTM model selectively keeps information over sequences using cell states and gates. The following describes the LSTM process.

First, the *forget gate* (f_t) evaluates the current input x_t with the output from the previous step h_{t-1}, applying a sigmoid function. The sigmoid function maps the input to a range between 0 and 1, deciding how much of each component the cell state should keep

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{1}$$

where W and b are the weights and biases of the gate, and σ is the sigmoid function. At the same time, the *input gate* chooses which new information to store in the cell state.

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{2}$$

After, the *cell gate* (g_t) generates a vector of new candidate values that could be added to the state, with the $tanh$ activation function, which outputs values between -1 and 1, allowing the model to regulate the updates to the cell state.

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{3}$$

The *output gate* then decides the LSTM's output at the time step, using the output of a sigmoid function to select values.

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{4}$$

and then the cell state is updated by combining the old state (multiplied by the *forget gate*'s output) and the new candidate values (multiplied by the *input gate*'s output)

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{5}$$

where $\odot$ denotes the Hadamard product. Finally, the hidden state for the current timestep is computed as

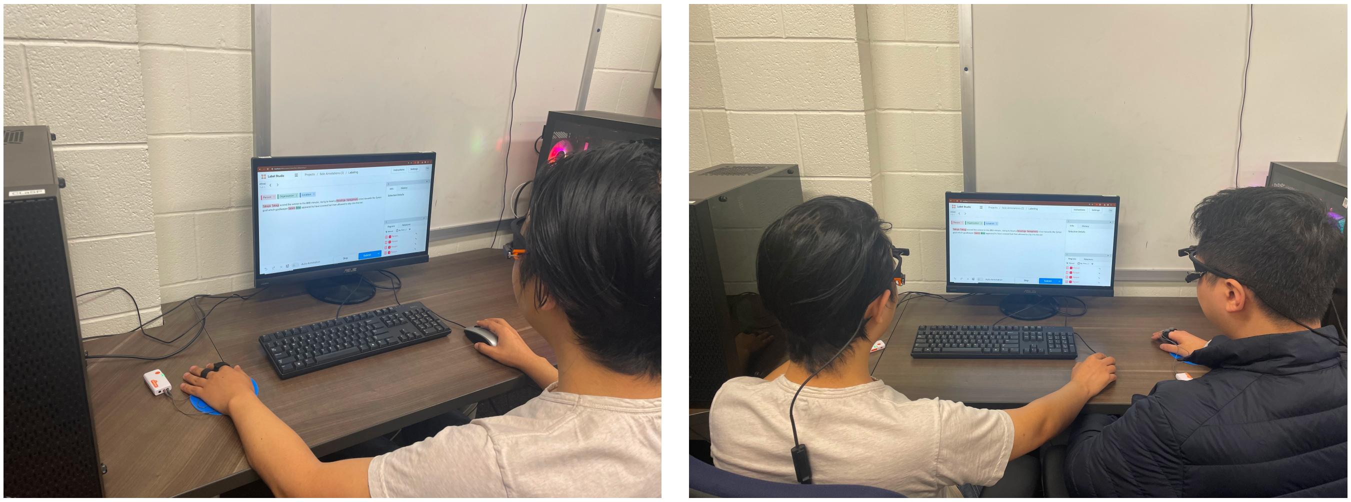
$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

The calculated hidden state h_t and the updated cell state C_t are then passed to the next timestep, completing the cycle of the LSTM. This process repeats for each element in the input sequence, which is important for the model to use previous information to make predictions, which is important in NER [54], [55].

The LSTM in this study was trained with the Adam optimizer with the following parameters (learning rate: 0.01, betas:(0.9, 0.999), eps:1e-08, weight decay:0) over 100 training epochs on an NVIDIA GeForce GTX 1070. The parameters were chosen to balance the LSTM model's learning capabilities with the necessity for human annotations, attempting to ensure the model benefited from human input without becoming overly accurate to the point of diminishing the value of human collaboration. The model's training was conducted on 14,000 samples from the CoNLL-2003 English dataset. The CoNLL-2003 dataset is a NER dataset in German and English introduced in 2003 by Sang et al [56]. The English portion of the dataset was taken from Reuters news stories between August 1996 and August 1997. The participant tasks were taken from the validation set of the CoNLL-2003 dataset.

This study's methodology uses a 2 × 2 factorial experimental design to examine the effects of two independent variables: pair-based collaboration (solo vs. pair) and system controllability (with vs. without) on cognitive load and model performance. This design allows for an understanding of the effects of the variables, giving insight into their impacts on user experiences and the quality of annotations in an interactive machine learning context. To achieve a statistical power of 80% a total of 128 participants will be targeted for recruitment so that 64 samples can be analyzed from each study (solo and pair) [57]. Participants are being recruited from an undergraduate psychology pool via the SONA system, which means the participants are predominantly students who are 18 to 22 years old. Participants can opt into the pair study to allow for within-subjects design.

Participants are tasked with annotating text data for NER, identifying and categorizing entities into three categories; person, location, and organization. The studies begin with an orientation where participants are familiarized with the NER interface and the equipment. Each participant undergoes a calibration phase for the eye tracker. Participants in the solo study complete the tasks individually (Figure 3a), while those in the pair study collaborate with a partner (Figure 3b). Each study is further divided into two sessions: one where participants annotate data without system controllability (Figure 2b) and another with controllability (Figure 2a). The order of these sessions is counterbalanced to mitigate the effect of order as shown in Figure 1. Each session has 20 NER annotation tasks, providing a consistent workload across participants and allowing for direct comparisons of cognitive load and annotation performance. The controllability aspect of



(a) Solo-based study where a single participant annotated NER tasks while wearing a GSR sensor and eye-tracker (capturing skin conductance and pupil diameter respectively).

(b) Pair-based study where two participants collaborate on the NER task while wearing a GSR sensor and eye-tracker (capturing skin conductance and pupil diameter respectively).

Fig. 3: Participants engaged in the solo-based and pair-based study setups.

this IML system is implemented through a browser extension where participants can change parameters affecting how the LSTM model was tuned on their annotations as seen in Figure 2a.

During these tasks, participants' biophysical responses are continuously monitored using the Shimmer3 GSR sensor and the Pupil Labs Core eye-tracking glasses (as seen in Figure 4). After each study session, participants are asked to complete the NASA-TLX rating scale survey. The NASA-TLX rating scale survey scores for each participant are calculated by averaging the participant's responses to the rating scale survey portion of the NASA-TLX survey. The average pupil dilation data for each participant was calculated by standardizing the pupil dilation values using z-score, smoothing them with a window size of 200 to reduce noise, and then taking the average of those values. Similarly, the average skin conductance for each participant was calculated by standardizing and smoothing the data with a window size of 200 and then taking the average.

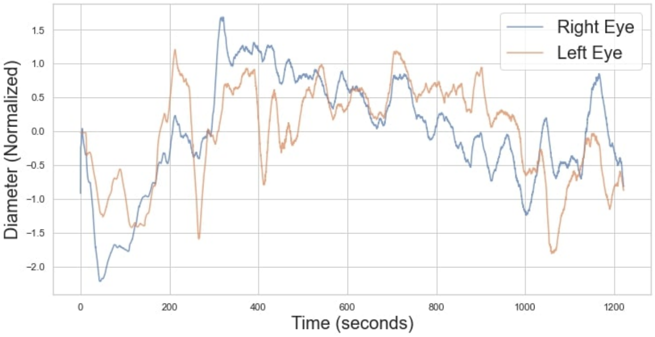


Fig. 4: Graphical representation of normalized and smoothed pupil diameters for a participant during solo annotation tasks with system controllability.

C. Progress

As of now, there have been 9 participants in the solo study and 6 participants in the pair study. This section reports the preliminary analysis of the current data. To analyze the data, a mix of independent and paired t-tests, Mann-Whitney U tests, Wilcoxon signed-rank tests, and mixed linear model regression analyses are being used to evaluate the effects of collaboration mode (solo vs. pair) and system controllability (with vs. without) on participant cognitive load and model performance, where model performance was measured by model's F1 score with respect to the CoNLL-2003 dataset. T-tests (independent and paired) compare mean values between two groups. The independent t-test is used when comparing two different groups (solo vs. pair). The paired t-test is used for comparing the same participants across two conditions (with and without controllability). Given the non-normal distribution of some of the data collected so far, such as physiological measures of cognitive load, the Mann-Whitney U and Wilcoxon signed-rank tests were used. The Mann-Whitney U test is a non-parametric test for comparing differences between two independent groups, and the Wilcoxon signed-rank test is a non-parametric test for comparing the same participants across two conditions. Mixed linear model regression analyses are also used to compare model performance because they account for both fixed effects (collaboration mode and system controllability) and random effects (variations among participants) in data with repeated measures.

In addressing RQ1, the preliminary results indicate that pair-based collaboration may not significantly affect cognitive load as measured by TLX scores, pupil diameter, or GSR. Specifically, the comparison of TLX scores between solo and pair modes did not reveal a significant difference (p-value: 0.76 for the independent t-test). Similarly, no significant difference was found in average pupil diameter (p-value: 0.07 for the Mann-Whitney U Test) and average skin conductance (p-value: 0.92 for the independent t-test). Additionally, the mixed linear model regression analysis on model F1 score showed no significant effect of collaboration mode when corrected for baseline model performance.

In addressing RQ2, statistically significant impacts were observed on cognitive load. The paired t-test for TLX scores between conditions with and without system controllability showed a notable increase in cognitive load for the controllable condition (p-value: 0.001 for the paired t-test). The average pupil diameter also indicated a notable increase in cognitive load for the controllable condition (p-value: 0.05 for the Wilcoxon Signed-Rank Test); however, average skin conductance did not (p-value: 0.68 for the paired t-test). The mixed linear model regression analysis for model F1 score showed no statically significant effect of the controllable condition when corrected for baseline model performance. These findings are preliminary and may change as additional data is collected and analyzed.

IV. Conclusion

Given the preliminary data, pair-based collaboration does not seem to significantly impact cognitive load or model performance in this IML task. Conversely, system controllability had a notable increase in cognitive load, showing the complexity and demands of giving users the ability to adjust system parameters.

The literature on cognitive load shows that in high cognitive load tasks, pair collaboration seems to reduce cognitive load more effectively than with simpler tasks [2]. Future works might see the effect of pair collaboration in IML systems with more intensive annotator tasks. Future work could also extend this investigation by incorporating additional analyses to further understand the subtleties of these interactions. Analyzing other metrics such as peak GSR and pupil diameter, heart rate variability, and more detailed pupillometry metrics could provide deeper insights into the physiological and psychological states of users interacting with IML systems along with a qualitative analysis of user preferences from feedback. The current study's limitations include the complexity and quantity of annotation tasks. Investigating the impact of pair collaboration in more demanding annotation tasks could give insights into its potential benefits in high cognitive load scenarios.

References

[1] E. Stefanidi, M. Bentvelzen, P. W. Woźniak, T. Kosch, M. P. Woźniak, T. Mildner, S. Schneegass, H. Müller, and J. Niess, "Literature reviews in HCI: A review of reviews," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, Apr. 2023, p. 1–24. [Online]. Available: https://dl.acm.org/doi/10.1145/3544548.3581332

[2] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*, ser. Volume 1. Springer New Yor, 2011. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4419-8126-4

[3] F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, B. Yin, and Y. Wang, "Multimodal behavior and interaction as indicators of cognitive load," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 4, pp. 22:1–22:36, Jan. 2013. [Online]. Available: https://dl.acm.org/doi/10.1145/2395123.2395127

[4] R. Pandey, H. Purohit, C. Castillo, and V. L. Shalin, "Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning," *International Journal of Human-Computer Studies*, vol. 160, p. 102772, Apr. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581922000015

[5] F. Westphal, N. Lavesson, and H. Grahn, "A case for guided machine learning," in *Machine Learning and Knowledge Extraction*. Cham: Springer International Publishing, 2019, pp. 353–361.

[6] A. G. Smith, J. Petersen, C. Terrones-Campos, A. K. Berthelsen, N. J. Forbes, S. Darkner, L. Specht, and I. R. Vogelius, "Rootpainter3D: Interactive-machine-learning enables rapid and accurate contouring for radiotherapy," *Medical Physics*, vol. 49, p. 461–473, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.15353

[7] M. H. Chignell, M.-H. Chung, Y. Yang, G. Cento, and A. Raman, "Human factors in interactive machine learning: A cybersecurity case study," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 1495–1499, 2021. [Online]. Available: https://doi.org/10.1177/1071181321651206

[8] M.-H. M. Chung, "Interactive machine learning in cybersecurity: Using human expertise more effectively," Ph.D. Thesis, University of Toronto (Canada), Canada – Ontario, CA, 2023. [Online]. Available: https://www.proquest.com/docview/2889776449/abstract/35CFF659674F447BPQ/1

[9] C. Alm and X. Llora, "Evolving emotional prosody," in *INTERSPEECH*, vol. 4, 09 2006.

[10] S. Koh, H. J. Wi, B. Hyung Kim, and S. Jo, "Personalizing the prediction: Interactive and interpretable machine learning," in *2019 16th International Conference on Ubiquitous Robots (UR)*, Jun. 2019, p. 354–359. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8768705

[11] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, Apr. 2023. [Online]. Available: https://doi.org/10.1186/s40537-023-00727-2

[12] A. Nandy, C. Duan, and H. J. Kulik, "Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery," *Current Opinion in Chemical Engineering*, vol. 36, p. 100778, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211339821001106

[13] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning: Solution and applications," *ACM Computing Surveys*, vol. 54, no. 10s, p. 1–29, Jan. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3502287

[14] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, p. 8934–8954, Sep. 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9941371?casa_token=9LMw-pxYy_YAAAAA:WMziaxbOBjUJIkcTYkcGSGHs02_z3ZaQxZ0GOJUu7uqs3Rt4T5blE6eQhWikSzw_bc78WC6j6RE

[15] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self- and unsupervised learning for image classification," *IEEE Access*, vol. 9, p. 82146–82168, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9442775

[16] M. Usama, J. Qadir, A. Raza, H. Arif, K.-l. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, p. 65579–65615, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8713992

[17] D. Valkenborg, A.-J. Rousseau, M. Geubbelmans, and T. Burzykowski, "Unsupervised learning," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 163, no. 6, p. 877–882, Jun. 2023. [Online]. Available: https://www.ajodo.org/article/S0889-5406(23)00193-2/fulltext

[18] K. Yoon and S. Kwek, "An unsupervised learning approach to resolving the data imbalanced issue in supervised learning

problems in functional genomics," in *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, Nov. 2005, pp. 6 pp.–. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1587765?casa_token=CxwdnboYmMMAAAAA:psOOBrcI8MLxu3-aDM_KKrbuLbFAySVfsqZ95yY2coyfQLE4iPUX_8nkUslsVAmlupa_VsPJimw

[19] A. Zimmermann, "Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 2, p. e1330, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1330

[20] R. Titung and C. Alm, "Teaching interactively to learn emotions in natural language," in *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 40–46. [Online]. Available: https://aclanthology.org/2022.hcinlp-1.6

[21] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, "Building trust in interactive machine learning via user contributed interpretable rules," in *27th International Conference on Intelligent User Interfaces*. Helsinki Finland: ACM, Mar. 2022, p. 537–548. [Online]. Available: https://dl.acm.org/doi/10.1145/3490099.3511111

[22] A. Tegen, P. Davidsson, and J. A. Persson, "A taxonomy of interactive online machine learning strategies," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, Sep. 2020, p. 137–153. [Online]. Available: https://doi.org/10.1007/978-3-030-67661-2_9

[23] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G. C. Crişan, C.-M. Pintea, and V. Palade, "Interactive machine learning: experimental evidence for the human in the algorithmic loop," *Applied Intelligence*, vol. 49, p. 2401–2414, Jul. 2019. [Online]. Available: https://doi.org/10.1007/s10489-018-1361-5

[24] F. Bernardo, M. Zbyszynski, R. Fiebrink, and M. Grierson, "Interactive machine learning for end-user innovation," in *AAAI Spring Symposia*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:14597497

[25] A. Tegen, P. Davidsson, and J. A. Persson, *Human Factors in Interactive Online Machine Learning*. IOS Press, 2023, p. 33–45. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/FAIA230073

[26] J. J. Dudley and P. O. Kristensson, "A review of user interface design for interactive machine learning," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 2, p. 1–37, Jun. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3185517

[27] M. Herde, D. Huseljic, B. Sick, and A. Calma, "A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification," *IEEE Access*, vol. 9, p. 166970–166989, Jan. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9650877/

[28] S. Mishra and J. M. Rzeszotarski, "Human expectations and perceptions of learning in machine teaching," in *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. Limassol Cyprus: ACM, Jun. 2023, p. 13–24. [Online]. Available: https://dl.acm.org/doi/10.1145/3565472.3595612

[29] J. Beck, "Quality aspects of annotated data," *AStA Wirtschafts- und Sozialstatistisches Archiv*, vol. 17, no. 3, p. 331–353, Dec. 2023. [Online]. Available: https://doi.org/10.1007/s11943-023-00332-y

[30] H. Takagi, "Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation," *Proceedings of the IEEE*, vol. 89, p. 1275–1296, Sep. 2001. [Online]. Available: https://ieeexplore.ieee.org/document/949485

[31] A. Singh, B. S. Minsker, and P. Bajcsy, "Image-based machine learning for reduction of user fatigue in an interactive model calibration system," *Journal of Computing in Civil Engineering*, vol. 24, no. 3, p. 241–251, May 2010. [Online]. Available: https://ascelibrary.org/doi/10.1061/%28ASCE%29CP.1943-5487.0000026

[32] V. J. Gawron, J. French, and D. Funke, *An Overview of Fatigue*. CRC Press, 2000.

[33] A. Fishbach and K. Woolley, "The structure of intrinsic motivation," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 9, no. Volume 9, 2022, pp. 339–363, 2022. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-orgpsych-012420-091122

[34] J. H. Jung, C. Schneider, and J. Valacich, "Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments," *Management Science*, vol. 56, no. 4, p. 724–742, 2010. [Online]. Available: https://www.jstor.org/stable/27784148

[35] C. C. Marshall and F. M. Shipman, "Experiences surveying the crowd: reflections on methods, participation, and reliability," in *Proceedings of the 5th Annual ACM Web Science Conference*, ser. WebSci '13. New York, NY, USA: Association for Computing Machinery, May 2013, p. 234–243. [Online]. Available: https://dl.acm.org/doi/10.1145/2464464.2464485

[36] M. Nadj, M. Knaeble, M. X. Li, and A. Maedche, "Power to the oracle? design principles for interactive labeling systems in machine learning," *KI - Künstliche Intelligenz*, vol. 34, no. 2, p. 131–142, Jun. 2020. [Online]. Available: https://doi.org/10.1007/s13218-020-00634-1

[37] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, p. 1–38, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988

[38] S. W. Kortschot, G. A. Jamieson, and A. Prasad, "Detecting and responding to information overload with an adaptive user interface," *Human Factors*, vol. 64, no. 4, p. 675–693, Jun. 2022. [Online]. Available: https://doi.org/10.1177/0018720820964343

[39] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_4

[40] B. G. D. S. Cezar and A. C. G. Maçada, "Cognitive overload, anxiety, cognitive fatigue, avoidance behavior and data literacy in big data environments," *Information Processing Management*, vol. 60, no. 6, p. 103482, 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306457323002194

[41] S. G. Hart and L. E. Staveland, *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, ser. Human Mental Workload. North-Holland, Jan. 1988, vol. 52, p. 139–183. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623869

[42] H. Devos, K. Gustafson, P. Ahmadnezhad, K. Liao, J. D. Mahnken, W. M. Brooks, and J. M. Burns, "Psychometric properties of nasa-tlx and index of cognitive activity as measures of cognitive workload in older adults," *Brain Sciences*, vol. 10, no. 12, 2020. [Online]. Available: https://www.mdpi.com/2076-3425/10/12/994

[43] S. Solhjoo, M. C. Haigney, E. McBee, J. J. G. van Merrienboer, L. Schuwirth, A. R. Artino, A. Battista, T. A. Ratcliffe, H. D. Lee, and S. J. Durning, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Scientific Reports*, vol. 9, p. 14668, Oct. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6789096/

[44] J. Yang, M. Tang, L. Cong, J. Sun, D. Guo, T. Zhang, K. Xiong, L. Wang, S. Cheng, J. Ma, and W. Hu, "Development and validation of an assessment index for quantifying cognitive task load in pilots under simulated flight conditions using heart rate variability and principal component analysis," *Ergonomics*, vol. 67, no. 4, p. 515–525, Apr. 2024. [Online]. Available: https://doi.org/10.1080/00140139.2023.2229075

[45] R. Bauer, L. Jost, B. Günther, and P. Jansen, "Pupillometry as a measure of cognitive load in mental rotation tasks with abstract and embodied figures," *Psychological Research*, vol. 86, no. 5, p. 1382–1396, Jul. 2022. [Online]. Available: https://doi.org/10.1007/s00426-021-01568-5

[46] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PloS One*, vol. 13, no. 9, p. e0203629, 2018.

[47] F. N. Biondi, B. Saberi, F. Graf, J. Cort, P. Pillai, and B. Balasingam, "Distracted worker: Using pupil size and blink rate to detect cognitive load during manufacturing tasks," *Applied Ergonomics*, vol. 106, p. 103867, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003687022001909

[48] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (gsr) as an index of cognitive load," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*. San Jose CA USA: ACM, Apr. 2007, p. 2651–2656. [Online]. Available: https://dl.acm.org/doi/10.1145/1240866.1241057

[49] N. Nourbakhsh, Y. Wang, and F. Chen, "Gsr and blink features for cognitive load classification," in *Human-Computer Interaction – INTERACT 2013*, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Berlin, Heidelberg: Springer, 2013, p. 159–166.

[50] P. K. Nihalani and D. H. Robinson, "Balancing collaboration and cognitive load to optimize individual and group desirable difficulties," *Journal of Educational Computing Research*, vol. 60, no. 2, p. 433–454, Apr. 2022. [Online]. Available: https://doi.org/10.1177/07356331211035188

[51] P. A. Kirschner, J. Sweller, F. Kirschner, and J. Zambrano R., "From cognitive load theory to collaborative cognitive load theory," *International Journal of Computer-Supported Collaborative Learning*, vol. 13, no. 2, p. 213–233, Jun. 2018. [Online]. Available: https://doi.org/10.1007/s11412-018-9277-y

[52] Z. Zhan, G. He, T. Li, L. He, and S. Xiang, "Effect of groups size on students' learning achievement, motivation, cognitive load, collaborative problem-solving quality, and in-class interaction in an introductory AI course," *Journal of Computer Assisted Learning*, vol. 38, no. 6, pp. 1807–1818, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12722

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, p. 1735–80, Dec. 1997.

[54] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, p. 357–370, Jul. 2016. [Online]. Available: https://doi.org/10.1162/tacl_a_00104

[55] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, no. arXiv:1402.1128, Feb. 2014, arXiv:1402.1128 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1402.1128

[56] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142–147. [Online]. Available: https://aclanthology.org/W03-0419

[57] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, no. 2, p. 175–191, May 2007. [Online]. Available: https://doi.org/10.3758/BF03193146

V. Appendix

I am currently a member of the Computational Linguistics and Speech Processing (CLaSP) lab. The CLaSP lab focuses on creating efficient, robust models that are human-centered and offer an IML framework. To further this goal, members of the ClaSP lab work on data analysis and annotation software, along with multi-modal ML models. My work focuses on the human factors in these models, ensuring human consideration while maintaining or improving model performance.