2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Toward Human-Like Assistance: Detecting Help-Seeking in Human–Robot Collaboration via Implicit Signals

Ane San Martin*, Ander Iriondo, Michael Hagenow, Julie Shah, Johan Kildal, and Elena Lazkano

***Abstract*— As collaborative robots become more common in industrial settings, enabling them to provide timely support to people when they are stuck is a key aspect. In particular, detecting when users need assistance, without relying on explicit requests, remains an open problem. This work explores whether machine learning models, such as Random Forest and Decision Tree, as well as the use of temporal dependencies can detect help-seeking behavior from implicit signals. Through a user study in a robot-assisted assembly task, we show that nonverbal cues, such as affective states and subtle behavioral dynamics, can reliably predict when a human needs assistance. Our top-performing model achieved an F1-score of 0.98. These findings demonstrate the feasibility of leveraging temporal modeling of implicit signals for proactive interaction in industrial contexts.**

I. INTRODUCTION

The presence of collaborative robots (cobots) in industrial environments is rapidly increasing due to their potential to improve flexibility and productivity [1]. However, achieving effective human–robot collaboration (HRC) requires more than physical coexistence; cobots must be capable of serving as a teammate, including offering assistance when the human worker needs help. This challenge becomes even more critical in industrial contexts where collocated human assistance is often absent. In such scenarios, workers may experience feelings of isolation and increased cognitive load as they struggle to solve issues on their own [2], [3]. These difficulties frequently result in delays in problem resolution, decrease in overall task productivity and trust [4], [5]. This reluctance has also been observed in the context of HRC, where recent findings highlight how hesitation to seek assistance shapes interaction dynamics [6], [7], [8].

Integrating AI-driven robotic assistance into these contexts offers a promising way to mitigate these issues. Compared to human assistance, AI and robots can provide non-judgmental support, reducing the stress and stigma associated with seeking help from others [9]. However, for robots to deliver proactive assistance, the first step is to identify when a user needs help without relying on explicit requests. In natural human–human collaboration, social signals such as gaze, posture, and facial expressions allow individuals to anticipate others' needs [10]. Inspired by this, we propose leveraging social signal processing to enable cobots to detect when users need help in a nonverbal, industrial context. These scenarios typically lack verbal interaction and mutual gaze, signals widely used in socially assistive robots for engagement and help detection, which makes adapting existing dialog-based approaches ineffective [11], [12].

While previous studies have demonstrated the value of using eye gaze to predict intention [13], [14] and facial Action Units (AUs) for detecting errors in robotic performance [15], these approaches have not yet been extended to predicting human help-seeking behavior in industrial scenarios. Identifying such moments is critical in industry, where timely support can prevent task breakdowns, reduce frustration, and improve collaboration efficiency. More important, focusing on help-seeking introduces another perspective compared to intention prediction or error detection, as it directly addresses the human's needs rather than focusing on the robot's performance. In this paper, we investigate the feasibility of detecting when users need assistance during industrial HRC using only implicit cues, such as gaze behavior, head movements, and affective states (valence and arousal), to identify help-seeking moments without relying on speech or explicit interaction. The main contributions of this work are threefold: (i) we train and evaluate machine learning (ML) models that predict when users need help using implicit cues collected during collaborative assembly tasks; (ii) we investigate how temporal dependencies affect the detection of help-seeking moments in industrial settings; and (iii) we investigate which variables are most informative for identifying situations in which a person needs help, and analyze how performance is affected through ablation studies of the most meaningful features.

II. RELATED WORK

A. Industrial Assistance in HRC

Given the increasing presence of robot teammates in industrial settings, much of the research in HRC has focused on detecting and correcting robot errors. For example, Stiber *et al.* [16] proposed a flexible error detection framework leveraging human responses, while Ravishankar *et al.* [17] introduced an online continual learning system for adaptive error prevention. Complementing these approaches, a hybrid AI framework inspired by Prognostics and Health Management (PHM) has been developed for fault detection in articulated cobots, enabling first-level maintenance by

*Corresponding author

A.S.M and J.S are with the Interactive Robotics Group, Massachusetts Institute of Technology, belonging to the Department of Aeronautics and Astronautics, Cambridge, Massachusetts, USA (anesm03@mit.edu, julie_a_shah@csail.mit.edu)

M.H is with the Department of Computer Sciences, University of Wisconsin - Madison, Madison, WI, USA (hagenow@cs.wisc.edu)

A.S.M, A.I and J.K are with the Department of Autonomous and Intelligent Systems, Tekniker, Eibar, Gipuzkoa, Spain ([ane.sanmartin|ander.iriondo|johan.kildal]@tekniker.es)

E.L is with the Faculty of Informatics, UPV/EHU, San Sebastian, Gipuzkoa, Spain (e.lazkano@ehu.eus)

alerting operators to anomalous trajectories [18]. In parallel, research on supporting human-in-the-loop error detection has explored how sharing the robot's knowledge base, through visualization or speech, can help users identify and correct errors during object organization tasks [19]. Together, these works illustrate the strong emphasis placed on efficiency-oriented approaches and error prevention in industrial collaboration, but they also reveal that comparatively little attention has been devoted to human needs, such as detecting when the human collaborator encounters difficulties, even in the absence of robot errors.

B. Implicit Signals and Machine Learning for Help Detection.

Inspired by human–human interaction, robotic research has begun to explore how social signals can inform adaptive robot behavior. For example, Tapus *et al.* [20] investigated physiological and affective indicators of cognitive load, using facial temperature signals to detect stress or malaise so that robots could adjust their support strategies in real time. Martin-Ozimek *et al.* [21] concentrated on multiparty social settings, developing methods to learn and interpret gaze patterns that enable robots to act as facilitators in group interactions. More aligned with our focus, Wilson *et al.* [11] advanced this line of research by proposing a multimodal architecture that combines gaze behavior with spoken language to infer when a user needs help, using ML models, triggering timely robot interventions. Earlier work by Reneau and Wilson [22] also highlighted the role of multimodal fusion, showing how integrating speech, gaze, and task models can improve a robot's ability to recognize when assistance is required and to respect user autonomy.

While prior work demonstrates the promise of social signal–driven assistance, much of it is situated in social robotics contexts where robots use expressive faces and spoken dialogue as primary communication channels. However, in industrial HRC, collaboration is task-focused and robots are typically limited to robotic arms without anthropomorphic features. Importantly, the absence of robot facial cues does not prevent the detection of help-seeking, since humans continue to display rich nonverbal signals when they require assistance. What it does mean is that assistance mechanisms cannot rely on face-to-face dialogue or expressive backchannels from the robot, but instead must interpret implicit human signals in real time. Our study addresses this gap by identifying nonverbal indicators that reliably signal when a human partner requires help, even in contexts where the robot itself lacks human-like expressivity. Unlike prior multimodal fusion frameworks that depend on explicit task models [22], our approach is task-agnostic, allowing it to generalize across diverse collaborative robotics settings. This adaptability is particularly valuable for industrial applications, where task variability and dynamic conditions demand flexible and robust assistance detection mechanisms.

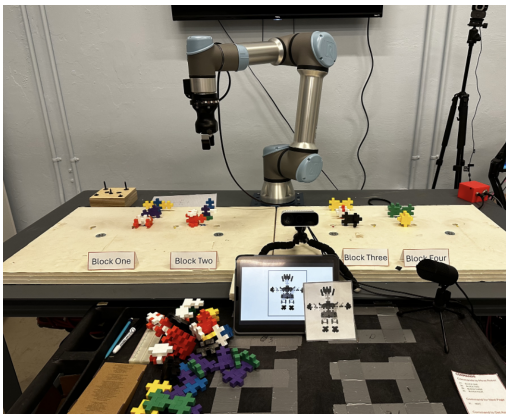

Fig. 1. Experimental setup used for dataset collection. The robot-assisted assembly task and components involved. The user was seated in front of the robot, some of the pieces were placed on the robot side and other in the human side. One camera was placed in front of the user to obtain facial information and another one in a corner for contextual information.

III. Methodology

In this work, we focus on detecting moments when users require assistance during HRC by analyzing implicit signals originating from the users, using data-driven ML algorithms. To generate a dataset suitable to train the models, we conducted a user study in which the need for help was deliberately elicited, ensuring the collection of representative instances of help-seeking behavior. This dataset forms the foundation for the subsequent preprocessing, temporal segmentation, and modeling steps described in the following sections.

A. Experiment Setup

The data was collected through a previous user study [8], where 20 participants were recruited (10 male, 10 female; ages 22–52, $M = 28, SD = 6.26$). The study was approved by the institutional review board.

In this study participants collaborated with a UR5 on a 3D puzzle assembly task (Fig. 1), following grayscale instructions displayed on a tablet to increase task difficulty. Some puzzle pieces were placed on the human's side, allowing direct access, while others were located on the robot's side. In those cases, the cobot was responsible for delivering the required pieces to the participant upon request. To ensure help-seeking, one critical piece was hidden mid-task. Interaction with the robot was handled through predefined voice commands (e.g., *Block One*), while explicit requests for assistance were made by saying "NEED HELP", which notified a remote assistant without visual access to the task.

The experimental setup included an Azure Kinect SDK camera capturing the participant, an additional camera positioned in the corner to record the overall scene, both capturing data at 15 fps, a microphone to communicate with the robot and the assistant, and two networked computers: one running *psi* for video and voice capture, and the other running ROS for robot control.

TABLE I
EXTRACTED STREAM FEATURES

Stream	Feature	Description
AUs	AU' Intensity	18 facial AUs [range 0-5]
Head	Head angles	3 features [roll, pitch, yaw] (radians)
Eyes	Gaze angles	2 features [pitch, yaw] (radians)
Facial Expression Recognition	valence and Arousal	2 features

B. Feature Extraction and Dataset Creation

For each experimental session, we extracted affective and behavioral features that are relevant for identifying help-seeking behavior [23]. These include facial AUs, head pose, and eye gaze, which provide fine-grained information about subtle expressions and attentional shifts, as well as emotional features such as valence and arousal. To compute these signals, frames were processed using two feature-extraction modules: OpenFace[1] and FaceChannel[2]. *OpenFace* is an open-source toolkit for facial behavior analysis that enables real-time estimation of facial landmarks, head pose, eye gaze, and 18 AUs [24], [25]. For each AU, it provides both occurrence (binary) and intensity (continuous, 0–5) values. In this work, we use only the intensity features. In parallel, *FaceChannel* is employed to derive valence and arousal measures from facial expressions. FaceChannel is a lightweight deep neural network specifically designed to reduce computational cost under dynamic conditions, incorporating an inhibitory layer that improves efficiency while lowering the number of trainable parameters [26].

The features extracted from these modules were combined to characterize users' affective and behavioral states and served as input for identifying moments where participants need help. To facilitate data collection and synchronization, all streams were integrated into Microsoft's Platform for Situated Intelligence (*psi*), which aligned the multimodal data at 333 ms intervals.

A summary of the features extracted using the aforementioned libraries and models, is detailed in Table I.

Once all the data were collected and stored, labeling was carried out in *psiStudio*[3] by two expert investigators, who independently annotated the recordings. The independent annotations showed an inter-rater agreement of 86.8%. Any disagreements were subsequently resolved through discussion until consensus was achieved. Following Csikszentmihalyi's concept of the *flow* state [27] and Rogers' framework of performance breakdowns [28], participant behavior was classified into four levels of assistance need:

- **Flow:** the participant is fully engaged in the task, showing no signs of difficulty or need for help.
- **Level 1:** the participant displays subtle hesitation cues (e.g., looking around, pausing, or shifting gaze), suggesting potential need for assistance.
- **Level 2:** the participant acknowledges the break in flow and attempts to resolve it using available resources (e.g., comparing the assembly with pictures), often alternating gaze between the structure and instructions.
- **Level 3:** the participant is unable to overcome the difficulty and explicitly requests help.

Since the goal of this work aims to determine whether models can be trained to identify when a user needs assistance and given the limited data available, levels 1–3 were grouped and labeled as *needing help* (label = 1), while flow instances were labeled as *not needing help* (label = 0). In total, the dataset comprises 75624 instances, corresponding to approximately 420 minutes of recorded data.

C. Data Preprocessing

Head and eye movement angles were then converted into frame-to-frame differences to capture changes over time rather than task-dependent absolute values. As a preliminary feature selection, 95% confidence intervals were analyzed over all features for "help" (1) vs. "no help" (0) instances. Features with non-overlapping intervals were selected, which yielded facial AUs (AU01, AU02, AU07, AU09, AU12, AU17, AU23, AU25, AU26, AU45), head movement differences (roll, pitch), gaze differences (yaw), and affective measures (valence, arousal). Interestingly, contrary to our expectations, users moved their heads less but shifted their gaze more when needing help.

All features were normalized to the range [0, 1] to ensure consistent scale and to improve model stability. To capture temporal dependencies, a sliding window approach was applied, either by (i) flattening the sequence into a single vector or (ii) averaging feature values within the window. Window sizes ranging from 1 to 50 (each sample representing 333 ms) were tested to examine how varying temporal contexts influenced model performance. Sliding windows have been widely used in sequential modeling, particularly for capturing dynamic patterns over time. For instance, Akkaladevi and Heindl [29] applied them in industrial action recognition to extract spatio-temporal descriptors efficiently, while Neto *et al.* [30] leveraged them in warehouse gesture-based interaction to improve robots' contextual understanding. Inspired by these applications, we adapted sliding windows to detect help-seeking behavior in HRC, enabling the models to incorporate temporal dynamics into their predictions.

Given the dataset imbalance (29% positive), both oversampling and downsampling were applied to obtain a 65/35 ratio between negative and positive classes. This ratio was chosen to mitigate bias toward the negative class while still reflecting the naturally lower frequency of help-seeking events observed in industrial contexts. After temporal processing, the dataset was shuffled to prevent ordering effects.

D. Training and Modeling

Model training was designed to assess how temporal context and feature selection affected the detection of help-

[1]https://github.com/TadasBaltrusaitis/OpenFace
[2]https://github.com/pablovin/FaceChannel
[3]https://github.com/microsoft/psi/wiki/Psi-Studio

TABLE II

MACHINE LEARNING MODELS AND PARAMETERS

Model	Key Parameters
Logistic Regression	max_iter = 1000
Decision Tree	max_depth = 15
SVM (RBF kernel)	default parameters
Random Forest	n_estimators = 1600, max_depth = 20
MLP	2 hidden layers (64, 32), ReLU, Adam, dropout = 0.2, max_epochs = 200

seeking behavior. We employed classical ML models (Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and SVM) together with a multi-layer perceptron (MLP). For training, a 5-fold cross-validation scheme was employed and repeated five times, ensuring that each subset served as a validation set once and reducing variance in performance estimates.

Hyperparameters for each model were chosen after iterative exploratory trials, where different configurations were tested and compared in terms of accuracy, stability, and computational cost. The final settings used are presented in Table II. These values provided a balance between predictive power and generalization without overfitting.

For each window size (1–50, 333 ms per step), the entire preprocessing pipeline: normalization, sliding window transformation, class balancing, and shuffling, was applied before training. This systematic setup ensured fair comparisons across models and temporal contexts.

To evaluate performance under class imbalance, three complementary metrics were used. The F1-score was computed for every model across all window sizes, while the ROC and Precision–Recall curves were examined for each model at the window size that achieved the highest F1-score. This combination provided a robust assessment of the models' ability to identify when users required assistance. Finally, for the best-performing model, SHAP analysis was conducted [31] to determine which features most strongly influenced predictions.

IV. RESULTS

Figures 2 and 3 show the F1-scores of the different models across various window sizes using the flatten and

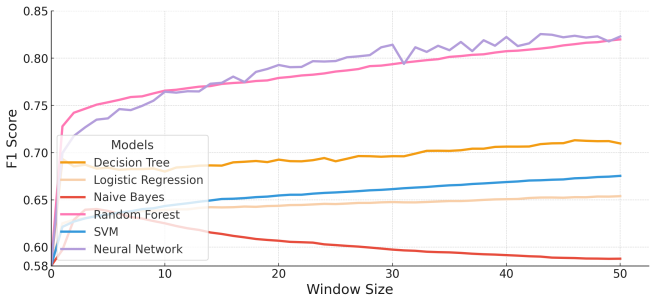

Fig. 2. F1 scores of all trained ML models across different window sizes, after flattening the features of the samples within each window.

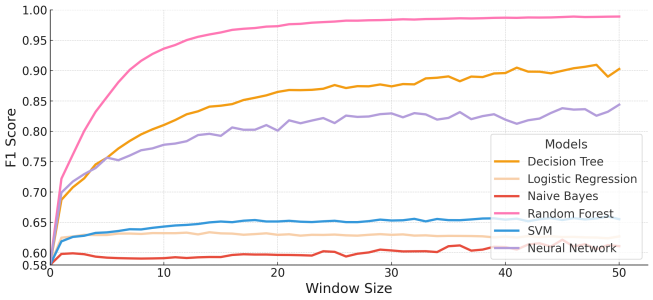

Fig. 3. F1 scores of all trained ML models across different window sizes, after averaging the features of the samples within each window.

mean methods, respectively. All reported values correspond to the average performance obtained through 5-fold cross-validation. As the results from downsampling and oversampling balancing strategies do not differ significantly, we focus on those obtained with the downsampling approach.

When using the flattening method on features within each window, we can observe that the models showing the best performance are Random Forest and MLP, with maximum F1 scores of 0.82 and 0.83. Moreover, the general trend across almost all algorithms is that performance improves as the window size increases (with the exception of Naive Bayes). However, we can see that even with a relatively small window, it is possible to achieve reasonably good performance, with only about a 6.1% loss.

However, by averaging the features of the different instances within a window, the best F1 scores are obtained. The Random Forest performs best with a window size of 40, achieving a maximum value of 0.98 and exceeding the best flattening-based result by 19.5%. In this case, an upward trend is also observed as the window size increases.

Although larger window sizes generally improved the performance of all models, this benefit comes at the cost of increased response time for the initial decision. Specifically, a longer temporal context requires the system to first accumulate sufficient observations, delaying the robot's ability to provide assistance until the buffer is filled. After this point, subsequent predictions proceed without additional delay. This trade-off underscores the need to balance predictive accuracy with responsiveness, particularly in industrial HRC scenarios where rapid adaptation is critical.

Among all the tested configurations, the best results were obtained when using the Random Forest classifier in combination with the mean function applied to the features within each temporal window. This approach not only yielded the highest F1-scores but also demonstrated robust generalization across different window sizes. In contrast, the MLP classifier performed competitively when trained with the full window, but its performance dropped considerably once the features were averaged within the windows. This suggests that simple models such as Random Forest are sufficient for this problem, as they already provide robust performance. To better understand the contribution of each feature to the Random Forest predictions, we conducted a SHAP analysis. This

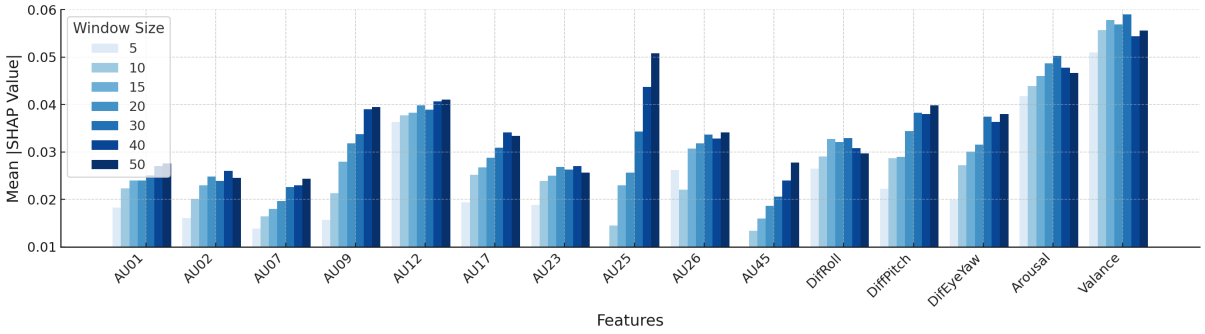

Fig. 4. SHAP analysis using the Random Forest algorithm across different temporal window sizes. A higher value indicates higher prediction power.

interpretability step allows us to identify the most influential features for detecting help-seeking behavior. Ultimately, the goal is to select a reduced subset of features that preserves most of the predictive power while improving computational efficiency, enabling faster execution and real-time deployment in collaborative industrial settings.

The results obtained in the SHAP analysis are depicted in Fig. 4. Among all variables, valence and arousal consistently showed the highest importance across all window sizes, confirming the relevance of affective states in detecting moments of difficulty. This aligns with prior findings suggesting that emotional cues such as frustration or uncertainty are strong indicators of a user's need for assistance [32], [33]. In addition to affective states, gaze shifts (DiffEyeYaw) and head movements (DiffRoll, DiffPitch) also contributed significantly to the model's decisions. These temporal dynamics capture hesitation and scanning behaviors, which often occur when participants struggle to continue with the task.

Facial AUs also showed notable contributions, with AU12 (lip corner puller), AU25 (lips part), AU17 (chin raiser), and AU09 (nose wrinkler) being the most relevant among them. These AUs likely reflect subtle expressions of confusion or effort. Other AUs, such as AU01 and AU02 (brow raisers), showed a smaller yet relevant influence, suggesting a role for fine-grained facial cues.

Guided by the SHAP analysis, we selected the three or four most influential variables for each window size. The logic behind this approach was to see whether a compact set of features could yield similar performance, enabling faster system response without a substantial drop in performance.

TABLE III

SELECTED FEATURE SUBSETS FOR TRAINING WITH 3 AND 4 VARIABLES ACROSS DIFFERENT WINDOW SIZES.

Window Size	Selected Features
4 Features	
5, 10, 15	Valence, Arousal, AU12, DiffRoll
20, 30	Valence, Arousal, AU12, DiffPitch
40	Valence, Arousal, AU25, AU12
50	Valence, AU25, Arousal, AU12
3 Features	
5, 10, 15, 20, 30	Valence, Arousal, AU12
40	Valence, Arousal, AU25
50	Valence, AU25, Arousal

Table III summarizes the feature combinations used for these experiments across different temporal window sizes.

Figures 5 and 6 present the Precision–Recall (PR) curves for the reduced feature configurations, using 3 and 4 features respectively. The 3 and 4 feature subsets achieved competitive results across all window sizes, though their performance was below that of the full feature set. Thus, when inference time and computational cost are not critical, using all features is preferable. Still, compact subsets such as 4 features can support reliable detection with only a modest drop in accuracy, making them valuable when very fast response times are required and a 5–12% trade-off is acceptable. Notably, the 4 feature set consistently showed higher precision at comparable recall levels across all window sizes, indicating a robust balance between efficiency and predictive power.

In contrast, the 3 feature configuration shows a systematic but moderate decrease in performance. Importantly, this reduction does not seem to depend strongly on the size of the temporal window: across short and long windows, the PR curves remain consistently below those of the four-feature case. This suggests that while valence, arousal, and a single AU (AU12 or AU25) are already highly informative, adding a fourth feature consistently improves the model's discriminative ability.

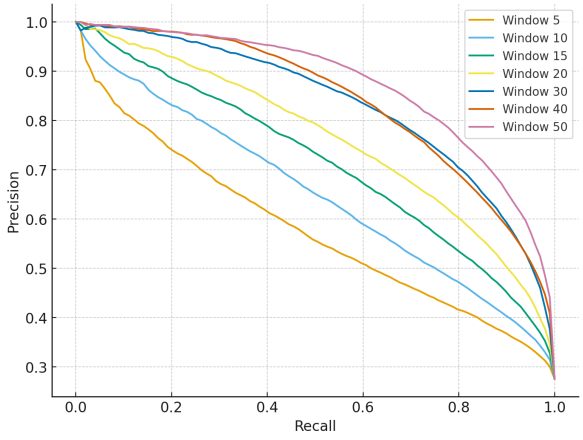

Fig. 5. PR curves using Random Forest and the best three-feature set for each temporal window.

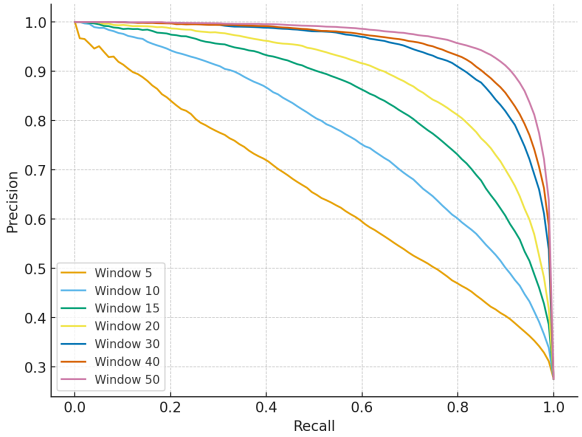

Fig. 6. PR curves using Random Forest and the best four-feature set for each temporal window.

Figure 7 compares the F1-scores obtained when training with three features, four features, and the full feature set across different temporal window sizes. As expected, models trained with all available features consistently achieve the highest F1-scores, with performance increasing steadily as the window size grows. This confirms that incorporating longer temporal contexts improves the model's ability to reliably detect help-seeking behavior. The reduced feature configurations show a systematic but moderate decrease in performance. Using four features yields notably better F1-scores than the three-feature setup for every window size, highlighting the contribution of the additional feature (typically a head movement dynamic such as DiffRoll or DiffPitch) in enhancing predictive accuracy. Importantly, the gap between the reduced sets and the full set remains relatively consistent across window sizes, indicating that the effect of feature reduction does not depend strongly on temporal context. Despite this decrease, both reduced configurations still achieve robust performance, particularly for larger windows (30–50), where F1-scores exceed 0.9 with four features and remain above 0.85 with three features. This demonstrates that a compact feature set can achieve near state-of-the-art results while significantly lowering computational cost, making real-time deployment more feasible.

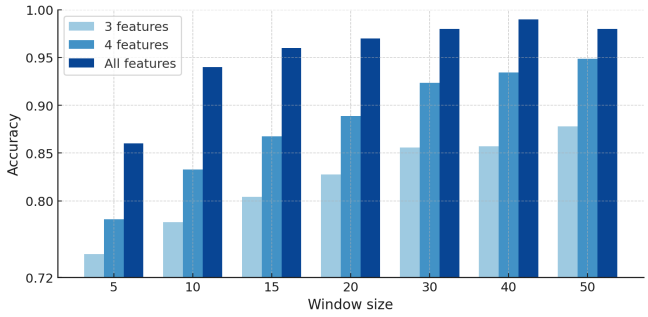

Fig. 7. F1-scores obtained using Random Forest and the best three/four-feature set for each temporal window.

V. Limitations

This study presents several limitations that should be acknowledged. First, the dataset was collected from a single controlled user study, which limits the generalizability of the findings. Although affective features and facial AUs appear to be task-independent indicators of help-seeking behavior, further validation across different HRC scenarios, task complexities, and user populations is necessary.

Second, the use of larger temporal window sizes improved predictive performance but introduced increased response latency. This creates a trade-off between detection accuracy and real-time responsiveness, which may limit practical deployment in time-critical industrial settings.

Finally, classical machine learning models were employed due to the relatively limited dataset size. While strong performance was achieved, the current approach may not fully capture complex temporal dependencies that could be better modeled with larger-scale data and more advanced sequential learning methods.

VI. Conclusions and Future Work

This work aimed to explore the viability of using ML models to detect moments when users require assistance during HRC, without relying on explicit task-related information. The results demonstrate that Random Forest, combined with temporal feature averaging and a window size of 40, yields the most effective detection performance, achieving F1-scores of up to 0.98. Moreover, the SHAP analysis revealed that affective variables, complemented by AU information and depending on the window, head dynamics, are the most informative indicators of help-seeking behavior. These findings confirm that it is possible to reliably infer user needs from multimodal behavioral signals, thereby reducing reliance on task-specific data. Notably, good F1 values can be achieved using only affective signals and AU information, which are largely independent of the specific task setup.

A key observation is the role of temporal context. While larger window sizes generally improved model accuracy, they also introduced longer response times, potentially limiting the robot's ability to respond on time. This underscores the trade-off between achieving high predictive accuracy and maintaining responsiveness in real-time support scenarios. Additionally, our results show that even small feature subsets (three or four variables) can deliver competitive performance, opening the door to more efficient real-time implementations. In our research, affective features and facial AUs emerged as reliable predictors of help-seeking behavior, and combining them with head-rotation dynamics provided additional discriminative power. This extends prior research on help recognition in HRC, which has often focused on voice or language-based cues [11] or gaze-language fusion [32], [34], which were not usually present in industrial settings.

Building on these findings, future work should focus on deploying the models in real-time environments to better identify the trade-off between predictive accuracy and the responsiveness required in HRC scenarios. Further studies across a broader range of tasks, collaboration contexts, and

user profiles will be essential to validate the robustness and general applicability of the identified behavioral indicators of help-seeking.

In addition, larger-scale datasets would enable the exploration of sequential deep learning approaches, such as Recurrent Neural Networks, which are well suited for modeling temporally dependent signals. Beyond binary detection, future research should investigate whether different levels or types of assistance need can be recognized, supporting more personalized robot interventions. Ultimately, extending these models toward proactive assistance strategies could allow robots not only to detect but also to anticipate user needs, leading to more adaptive support, improved task efficiency, and enhanced user experience in human–robot collaboration.

ACKNOWLEDGMENT

The research for this paper has been financially supported by the Elkartek Programme, Basque Government (Spain), project OSADAPT KK-2025/00028.

REFERENCES

[1] L. F. Concetta Manuela, G. Antonio, L. S. Giada, M. Rosa *et al.*, "Toward acceptance of human-robot collaboration in industrial settings: a bibliometric and systematic literature review," *The International Journal of Advanced Manufacturing Technology*, pp. 1–22, 2025.

[2] F. J. Milliken, E. W. Morrison, and P. F. Hewlin, "An exploratory study of employee silence: Issues that employees don't communicate upward and why," *Journal of management studies*, vol. 40, no. 6, pp. 1453–1476, 2003.

[3] J. Singh, "Performance productivity and quality of frontline employees in service organizations," *Journal of marketing*, vol. 64, no. 2, pp. 15–34, 2000.

[4] R. Ferreira, R. Pereira, I. S. Bianchi, and M. M. da Silva, "Decision factors for remote work adoption: advantages, disadvantages, driving forces and challenges," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 7, no. 1, p. 70, 2021.

[5] T. D. Golden, J. F. Veiga, and R. N. Dino, "The impact of professional isolation on teleworker job performance and turnover intentions: does time spent teleworking, interacting face-to-face, or having access to communication-enhancing technology matter?" *Journal of applied psychology*, vol. 93, no. 6, p. 1412, 2008.

[6] V. K. Bohns and F. J. Flynn, ""why didn't you just ask?" underestimating the discomfort of help-seeking," *Journal of Experimental social psychology*, vol. 46, no. 2, pp. 402–409, 2010.

[7] F. Lee, "When the going gets tough, do the tough ask for help? help seeking and power motivation in organizations," *Organizational behavior and human decision processes*, vol. 72, no. 3, pp. 336–363, 1997.

[8] Title and authors omitted for anonymous review, "Title and authors omitted for anonymous review," *arXiv preprint*, 2025.

[9] K. Kavitha, V. Joshith, and S. Sharma, "Beyond text: ChatGPT as an emotional resilience support tool for gen z–a sequential explanatory design exploration," *E-Learning and Digital Media*, p. 20427530241259099, 2024.

[10] N. Hulle, S. Aroca-Ouellette, A. J. Ries, J. Brawer, K. Von Der Wense, and A. Roncone, "Eyes on the game: Deciphering implicit human signals to infer human proficiency, trust, and intent," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2024, pp. 453–460.

[11] J. R. Wilson, P. T. Aung, and I. Boucher, "When to help? a multimodal architecture for recognizing when a user needs help from a social robot," in *International Conference on Social Robotics*. Springer, 2022, pp. 253–266.

[12] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: in-the-wild experiences and lessons learned," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 637–644.

[13] B. Yang, J. Huang, X. Chen, X. Li, and Y. Hasegawa, "Natural grasp intention recognition based on gaze in human-robot interaction," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 2059–2070, 2023.

[14] H. Zheng, W. Hongxing, Z. Tianpei, and Y. Bin, "The collaborative power inspection task allocation method of "unmanned aerial vehicle and operating vehicle"," *IEEE Access*, vol. 9, pp. 62 926–62 934, 2021.

[15] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 222–230.

[16] M. Stiber, "Flexible robot error detection using natural human responses for effective hri," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 148–150.

[17] J. Ravishankar, M. Doering, and T. Kanda, "Online continual learning for automatic error prevention in data-driven hri," *ACM Transactions on Human-Robot Interaction*, 2025.

[18] A. Polenghi, L. Cattaneo, and M. Macchi, "A framework for fault detection and diagnostics of articulated collaborative robots based on hybrid series modelling of artificial intelligence algorithms," *Journal of Intelligent Manufacturing*, vol. 35, no. 5, pp. 1929–1947, 2024.

[19] H. A. Frijns, M. Hirschmanner, B. Sienkiewicz, P. Hönig, B. Indurkhya, and M. Vincze, "Human-in-the-loop error detection in an object organization task with a social robot," *Frontiers in Robotics and AI*, vol. 11, p. 1356827, 2024.

[20] A. Tapus *et al.*, "Robots detecting cognitive load (e.g., stress) using facial temperature signals for adaptive support," 2025, eNSTA Paris / Institut Polytechnique de Paris news article, RAICAM project.

[21] A. Martin-Ozimek, I. Jayarathne, S. L. Mon, and J. Chew, "Learning nonverbal cues in multiparty social interactions for robotic facilitators," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 1042–1046.

[22] A. Reneau and J. R. Wilson, "Supporting user autonomy with multimodal fusion to detect when a user needs assistance from a social robot," *arXiv preprint arXiv:2012.04078*, 2020.

[23] G.-Y. Wang, Y. Hatori, Y. Sato, C.-H. Tseng, and S. Shioiri, "Predicting learners' engagement and help-seeking behaviors in an e-learning environment by using facial and head pose features," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100387, 2025.

[24] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[25] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[26] P. Barros, N. Churamani, and A. Sciutti, "The facechannel: a fast and furious deep neural network for facial expression recognition," *SN Computer Science*, vol. 1, no. 6, p. 321, 2020.

[27] M. Czikszentmihalyi, *Flow: The psychology of optimal experience*. New York: Harper & Row, 1990.

[28] J. Rogers, "Performance assessment of self-care skills," *Rehabilitation Psychology*, 1989.

[29] S. C. Akkaladevi and C. Heindl, "Action recognition for human robot interaction in industrial applications," in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*. IEEE, 2015, pp. 94–99.

[30] P. Neto, M. Simão, N. Mendes, and M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 101, pp. 119–135, 2019.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[32] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[33] J. Morriss, E. Tupitsa, H. F. Dodd, and C. R. Hirsch, "Uncertainty makes me emotional: Uncertainty as an elicitor and modulator of emotional states," *Frontiers in psychology*, vol. 13, p. 777025, 2022.

[34] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.