# Personalization of Industrial Human-Robot Communication through Domain Adaptation based on User Feedback

**Debasmita Mukherjee**
University of British Columbia
**Jayden Hong**
University of Victoria
**Haripriya Vats**
Indira Gandhi Delhi Technical University for Women
**Sooyeon Bae**
University of Toronto
**Homayoun Najjaran**
najjaran@uvic.ca

University of Victoria

**Research Article**

**Keywords:** human-robot collaboration, human-robot communication, multimodal communication, personalized machine learning, human feedback, robot perception, facial expression recognition, model personalization

**Posted Date:** March 9th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2656781/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at User Modeling and User-Adapted Interaction on March 22nd, 2024. See the published version at https://doi.org/10.1007/s11257-024-09394-1.

# Personalization of Industrial Human-Robot Communication through Domain Adaptation based on User Feedback

Debasmita Mukherjee[1], Jayden Hong[2], Haripriya Vats[3], Sooyeon Bae[4] and Homayoun Najjaran[2*]

[1]School of Engineering, The University of British Columbia, Alumni Avenue, Kelowna, V1V 1V7, British Columbia, Canada.
[2]Faculty of Engineering and Computer Science, University of Victoria, Finnerty Road, Victoria, V8P 5C2, British Columbia, Canada.
[3]Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, Madrasa Road, Delhi, 110006, Delhi, India.
[4]Faculty of Applied Science and Engineering, University of Toronto, Street, Toronto, 610101, State, Canada.

*Corresponding author(s). E-mail(s): najjaran@uvic.ca;
Contributing authors: debasmita.mukherjee@alumni.ubc.ca; jaydenh@uvic.ca; haripriyavats@gmail.com; sooyeon.bae@mail.utoronto.ca;

**Abstract**

Achieving safe collaboration between humans and robots in an industrial work-cell requires effective communication. This can be achieved through a robot perception system developed using data-driven machine learning. The challenge for human-robot communication is the availability of extensive, labelled datasets for training. Due to the variations in human behaviour and the impact of environmental conditions on the performance of perception models, models trained on standard, publicly available datasets fail to generalize well to domain and application-specific scenarios. Thus, model personalization involving the adaptation of such models to the individual humans involved in the task in the given environment would lead to better model performance. A novel framework

is presented that leverages robust modes of communication and gathers feedback from the human partner to auto-label the mode with the sparse dataset. The strength of the contribution lies in using in-commensurable multimodes of inputs for personalizing models with user-specific data. The personalization through feedback-enabled human-robot communication (PF-HRCom) framework is implemented on the use of facial expression recognition as a safety feature to ensure that the human partner is engaged in the collaborative task with the robot. Additionally, PF-HRCom has been applied to a real-time human-robot handover task with a robotic manipulator. The perception module of the manipulator adapts to the user's facial expressions and personalizes the model using feedback. Having said that, the framework is applicable to other combinations of multimodal inputs in human-robot collaboration applications.

**Keywords:** human-robot collaboration; human-robot communication; multimodal communication; personalized machine learning; human feedback; robot perception; facial expression recognition; model personalization

# 1 Introduction

Researchers envisage an industrial work-cell in which in the truest sense of collaboration, the human partner contributes their versatility, precision, and dexterity in carrying out tasks while the robot partner tackles repetitive, non-ergonomic and physically taxing tasks. Named aptly, the field of human-robot collaboration (HRC), aims to bring in a dramatic transformation in manufacturing. In order to carry out tasks in dynamic environments such as in industries, the robot partner, much like the human one must be aware of its environment, the status of the human and the shared task. This understanding of the robot's environment can be tackled through effective communication between the human and robot partners.

Perception is the process of selection, organisation, and categorization or interpretation of data to understand the environment or the agent's internal state. Sorting and organising of information is carried out through learnt cognitive patterns. While perception for humans is a psychological and cognitive process, the outcome of the operations influence human communication. For robots collaborating with humans too, the first step to communication is robot perception.

With the aim to develop a communication system for collaboration that is as close to "natural" communication as possible, researchers are looking to incorporate multiple modes of input from the human partner into the robot perception system. These multimodal inputs include voluntary modes such as hand gestures and voice commands and involuntary ones such as facial expressions, gaze, and body language (Mukherjee et al. 2022a; Skantze et al. 2014). The involuntary modes are indicative of the internal state of the humans (Ekman and Friesen 2003) while the voluntary modes can be used in the

imperative sense. Thus, a typical HRC system in an industrial setting that serves to mimic the ease with which humans collaborate must necessarily allow for robot perception the means to understand both lines of communication in order to ensure both physical as well as psychological safety of the involved human partner.

Currently, multimodal inputs are detected and classified by machine learning (ML) models by learning a generic feature representation using a large dataset such as Imagenet (Russakovsky et al. 2015) or MS COCO (Lin et al. 2014) with over a million labelled images and then fine-tuning based on the required specific application. This framework is followed for almost all aspects of supervised learning such as semantic segmentation (Long et al. 2014), object detection (Girshick et al. 2013) and pose estimation (Tulsiani and Malik 2014). Neural network architectures have been developed to leverage these features (Krizhevsky et al. 2012),(Simonyan and Zisserman 2014),(Chatfield et al. 2014) and achieve high accuracies but with the bottleneck that these require extensive datasets for learning.

The fulfillment of requirement for large, labelled datasets proves challenging for HRC applications. Human behaviour in its natural, intuitive state may be highly personalized, and while industrial commands may be "standardised", some aspects may still be atypical and idiosyncratic to the human partner. For example, from a human's visual perspective, say a hand gesture in a variety of settings, under different lighting conditions, from different gesturing humans with their characteristic ways may be similar, but to the robot perception system, will invariably lead to domain shifts in the incoming data. Most existing ML models are trained with the assumption of being operated in a closed-world scenario i.e., the test data is drawn from the same distribution as the training data. Such a case is known as the data being in-distribution (ID). However, on being deployed in the real-world, with the dynamics and uncertainties attached to the continually changing environment, test data are often sourced from a distribution different than that of the source. Such shifted data lead to poor performance of the model, even though the latter may be well-trained on ID data. Instead of rejecting these observations, the perception model must be updated. This is especially significant for recognizing affect data such as emotions through facial expressions and body language. Hence, the models must be "personalized" to the behaviour of the human partner for a more humanistic design of industrial systems.

On that note, in order to ensure the safety of the human partner during an HRC task, the former's focus on the task gauged through "positive" communication signals must be continuously gathered and recognized. A particular scenario has been examined in this paper built upon the general framework presented in a previous work (Mukherjee et al. 2022b) particularly on whether the human partner is "engaged" in the task. Here "engaged" refers to the lexical meaning of having one's interest or attention occupied. Emotions through facial expressions are utilized to provide an indication of the focus of the human partner on the collaborative task. In an online operation, the robot would read

the human's facial expressions and if the human is found to be expressing "not engaged" expression perhaps due to surprise, anger, fear or confusion, the robot would pause operation and wait for further instruction in order to ensure safety. On the other hand, if the expression was found to be "engaged" in the task i.e., showing signs of "happiness" or simply being "neutral", then the robot operation would continue. In this way, the affective state of the human would be taken into consideration to ensure their complete focus and safety.

The key difference however between facial expression recognition (FER) and other modes such as voice commands and hand gestures is the lack of large publicly available annotated datasets of the former. Voice commands can be learned using the Google Speech Commands dataset (Warden 2018) (65,000 seconds of recorded and annotated data) or Mozilla's Common Speech dataset (Mozilla 2022) (more than 14,000 hours of validated data in 93 languages). FER is learned on datasets such as which are based on socio-affective human behaviours. Some publicly available ones are AffectNet (440,000 annotated images sourced from the internet through search engine querying) (Mollahosseini et al. 2019) , CK+ (Extended Cohn-Kanade dataset) (593 video sequences under laboratory conditions) (Lucey et al. 2010), FER2013 (30,000 images) (Goodfellow et al. 2013) and JAFFE (213 images of only Japanese female participants) (Zhao and Zhang 2011). Most of the images portray exaggerated expressions that are a consequence of being acted out and may not be representative of "real-world" expressions in a given setting. Recent work in ascertaining the labelling quality of AffectNet has shown that only 13% manual votes retained the label of images from a portion of the original AffectNet dataset (Kim and Wallraven 2021). Indeed, participants in the aforementioned study preferred to label images as "neutral" expression without knowledge of the context of the expression. This highlights a second challenge of using affective data – the variability of expression for a given emotion in humans (Barrett et al. 2019). Naturally, models trained with limited, lab-recorded data would fail to capture the diversity of humans' expressions without significantly expanding the dataset. Studies have corroborated this limitation wherein FER in real-time generated lower accuracies as compared to pre-defined datasets (Rawal and Stock-Homburg 2022). Even if the hefty endeavour of collecting and labelling individualized data were to be taken, labelling expressions out of context may end up with an unusable dataset that could be detrimental to the safety of the human working with the robot. Finally, the vagaries of expressions in humans are coupled with the change in lighting conditions, background, properties of the data (dimensions, relative positioning of the human from the robot) that trip up computer vision models.

In order to deploy an industry-ready perception model that recognizes expressions with sufficient accuracy, the models trained with standard datasets need to be re-trained on the job with application-specific ones. These datasets involving humans will naturally be much more sparse as compared to the publicly available ones. Thus the emphasis of the proposed framework in this paper

is on utilising as small batches of user data as possible to generate personalized models that can ensure accurate recognition of the user data.

# 2 Related Works

## 2.1 Related Human-Robot Communication Concepts

**Facial Expression Recognition in HRC**:
The recognition of human emotions through their facial expressions has been studied for human-robot communication (HRCom) in order to endow robots with emotional intelligence. This is due to human emotions having a significant impact on their decisions and actions. The reader is directed towards the review covering FER and other modes of emotion recognition (Spezialetti et al. 2020), emotion recognition for HRC (Mohammed and Hassan 2020), and FER applied to human-robot interaction (Mukherjee et al. 2022a; Rawal and Stock-Homburg 2022). A facial expression emotion recognition-based human-robot interaction (FEER-HRI) system was proposed. It enabled the robot to recognize human emotions and then generate facial expressions using symbols on an LED screen to respond to them (Liu et al. 2017). Images from JAFFE dataset were preprocessed using face detection and segmentation. Regions of interest were created and features were extracted. FER was carried out using 2D-Gabor, uniform local binary pattern operator, and multi-class extreme learning machine (ELM) classifier. Similarly, in (Hsu et al. 2017), Gabor filters followed by support vector machine was used to obtain action units which were applied to random forest classifiers to recognize facial expressions for an interaction task. KDEF dataset (Kale et al. 2022) was used in an interaction scenario where the robot responded based on the recognized emotion of the human (Faria et al. 2017). Dynamic Bayesian Mixture Model and feature-based ML classifiers were employed to detect and recognize affective facial expressions. The highest accuracy achieved was around 85% on KDEF and 80% on data collected from human participants involved in the study. While there has been much success with traditional ML algorithms, the field of FER is increasingly employing deep learning algorithms that involve lesser preprocessing of data (Rawal and Stock-Homburg 2022).
**Multimodal Communication in HRC**:
A multimodal emotional communication-based human-robot interaction system was presented (Liu et al. 2017). It consisted of cameras to record real-time images of facial expressions and body gestures, a microphone to capture speech signals, and an eye tracker to study interaction in four scenarios: guiding, entertainment, home service, and scene simulation. Other multimodes such as hand gestures and speech are also widely used as communication for HRC operations. Pointing gestures and voice commands are utilized to provide information for the processing of requests from the robot (Maurtua et al. 2016,1). The information from the two modes is fused. Contradictory information is handled by semantic technologies to construct a command for the robot. Directional hand gestures and corresponding voice commands were

used to guide a human-robot handover task in (Mukherjee et al. 2022b). The fusion architecture was designed using a fuzzy inference system and Dempster Shafer theory to handle conflicting, complementary, and ambiguous commands from multimodal communication. Simulations in virtual reality were executed to design a safe human-robot collaborative nut-screwing task (Shu et al. 2019). Multimodal inputs were gathered through a GUI for examining use cases and identifying the most efficient manner of communication for the task. Additionally, multimodal inputs have been developed in tandem using fusion systems (Rossi et al. 2013), (Liu et al. 2018), (Reddy and Basir 2010) as well as individually (Chen et al. 2018; Drawdy and Yanik 2015; Nuzzi et al. 2021; Rautiainen et al. 2022; Wang et al. 2019).

**Feedback in HRC**:

Taking a cue from human communication, feedback serves as a powerful tool to provide guidance to the collaborative robot in a variety of tasks. Communication in a collaborative scenario should include a means of feedback from the human partner regarding the status of the task, deviations in the expected behaviour of the robot, and the general condition of the human in the work-cell (Kardos et al. 2018). In terms of the usage of human-on-the-loop feedback in HRC scenario, the work cited in (Wilde et al. 2018) presented robot path planning using human feedback. The methodology was used to learn the human's preferences for spatial and simple temporal constraints in solving a shortest-path problem. Feedback was gathered by presenting the user with alternate paths. In the field of affective computing, an approach was developed for learning user preferences of the robotic sculpture's motion during on-line human-robot interaction (Kumagai et al. 2018). The human partner's facial expressions were detected and recognized during the interaction based on which reward function was formulated. The robot was rewarded when the human's positive affect was observed through their expressions. This led to adapting the action parameters to maximise the reward function during reinforcement learning. This allowed the system to be customized to the users' preferences.

## 2.2 Personalization of ML models for HRC

Outside of HRC, personalization of ML models through the use of integrated, individualized patient datasets and application of ML algorithms in clinical workflows are increasingly being used to inform better healthcare e.g., in the dermatology (Wongvibulsin et al. 2022), determining which candidates would be suitable for cognitive training (Shani et al. 2021), and spine care (Khan et al. 2020). In the realm of human-computer interaction, speech emotion recognition (SER) using personal voice data was studied in (Kim and Park 2016). Personalized SER was achieved through adaptation using maximum likelihood linear regression. It was developed utilizing voice data collected from personal handheld communication devices such as smartphones. The approach selected useful emotionally discriminative acoustic characteristics. An iterative

unsupervised scheme was developed for the automatic labelling of SER data leading to higher accuracy and reduced recognition errors. Intelligent agent technologies, also known as chatbots were examined through 57,000 interactions to determine if human-computer interactions could be more personalized by matching the inferred personality of the human to the learnt personality of the machine. This matching of personalities had a positive impact on consumer engagement and purchasing (Shumanov and Johnson 2021).

Active learning based on computer vision was used to recognize humans, profile them and then personalize the robot behaviour. Humans were identified using Intel-face-detection-retail-004 and FaceNet for face recognition and other information was obtained using interactions with the robot (Maroto-Gómez et al. 2023). Indeed, several studies have demonstrated that adaptation and personalization of robot perception systems and in turn behaviour to that of the human improves the quality of interaction and leads to greater user acceptance (Caleb-Solly et al. 2018; Churamani et al. 2017; Di Napoli et al. 2018). Apart from that, researchers found that gesture personalization during a collaborative task reduced the mental and physical workload of the humans and was thus increasingly preferred by the participants (Rautiainen et al. 2022).

Emotion recognition is being used to check driver impairment and personalize the driving experience by numerous companies like Affectiva (Affectiva 2018). From a research perspective, reliable driver state recognition was studied in (Yi et al. 2019) as a precursor to driver safety monitoring systems and adaptive driving assistive systems. A personalized driving state recognition system dedicated to individual drivers was developed for higher accuracy of state (normal, drowsy, and aggressive) recognition, and expected improvement in road safety. Individual drivers' feature data were analysed and compared to generic models. If significant differences were detected, personalized models were developed for predicting those drivers' states. While the study demonstrated improved accuracy as compared to using the generic model, the dataset used was manually labelled and learning was carried out offline. For continuous adaptation and personalization in safety-critical operations, there is a need for the inclusion of online operation of such algorithms as well as a means for automatic labelling.

## 2.3 Adaptation of Shifted Data

Re-training of models for adapting to shifted data has been explored extensively using knowledge distillation and more particularly cross-modal knowledge distillation. A model updating framework was presented based on lifelong ML to counter calibration drifts in prediction models using soft labels for knowledge distillation (Chi et al. 2022). Researchers developed a scheme named 'supervision transfer' in which a large set of unlabeled paired RGB and depth images of the same objects (NYUD2 dataset) were used for transferring the ImageNet supervision on RGB images to depth images (Gupta et al. 2015). A key aspect of the scheme developed is that the modalities were "paired" i.e.,

they represent the same object(s) to be detected within the same setting and with the lighting conditions consistent throughout. Also, a major assumption was the accessibility of a large unannotated dataset containing paired data. In a similar vein, in (Wang et al. 2021), cross-modal knowledge distillation for Cued Speech (CS) was presented which consisted of lip movements along with synchronized hand movements. In order to tackle the limited size of CS data, researchers developed a novel system to transfer and preserve knowledge across the modalities. A large amount of open-source audio speech data was used to pre-train a teacher model. Then the speech knowledge was distilled into the small student model through two strategies- frame-level and sequence-level. Although the two modalities considered are from audio and video sources, the two modalities contain the same phoneme semantics due to both being synchronous and phoneme-level coding. Synchronised RF signals and camera images were used to transfer the knowledge across modalities for pose estimation in (Zhao et al. 2018) and in (Thoker and Gall 2019), the teacher model was trained on RGB videos and the student model on 3D human pose sequences paired with the videos.

Another approach that has been gaining ground is the Domain Adaptive Knowledge Distillation method developed in (Kothandaraman et al. 2020) contributes to the field of Unsupervised Domain Adaptation (UDA) wherein the goal is to align features and reduce the gap between the labelled source domain and the unlabelled target domains to boost model performance in both domains. The methodology was developed on RGB images in both the source as well as target domains. Thus, while literature exists for "paired" sources of data, more innovative methodologies must be looked into in fields such as FER wherein such extensive and synchronized labelled data is not available. Not only that, multimodal data can often be incommensurable such as in the case of facial expressions and voice commands. In such cases, an off-the-shelf application of knowledge distillation or domain adaptation techniques may not be feasible. The framework proposed in this paper thus takes inspiration from UDA and uses transfer learning as a means of adapting and personalizing the FER model to an industrial collaborative task.

## 3 Contributions

In order to develop a personalized ML system, one must take advantage of the existence of publicly available large datasets that can provide feedback during continuous operation so that new data from the user can be labelled using that feedback. This mimics a key component of natural human communication wherein feedback is used to continually update the knowledge and status associated with the shared task (Mukherjee et al. 2022c). The present document proposes a framework for such a continuous re-training leading to the development of personalized ML models for HRCom. Personalization is to be achieved for the environment and the human collaborating with the robot, thus requiring greater specificity than a generic model trained on publicly available standard datasets.

The specific application considered is the FER system that is trained with a standard publicly available dataset and then iteratively re-trained with users' images with both clean, uncluttered background and also with more complex, cluttered background of the same subject(s). Since these are unlabelled data and a fraction of size of the original dataset (which itself is typically a small dataset), the feedback from voice commands from the user is used to assign labels to these data. Once trained with the data of the user the robot or more generally, the system would encounter in its collaborative work, the re-training would slow down until new users are added to the work-cell. The personalization through feedback-enabled human-robot communication (PF-HRCom) framework has been developed in Section 4 with ablation tests in Section 6. Finally, PF-HRCom has been applied to a real-time human-robot handover task between a 6 degree of freedom (DoF) manipulator: Kinova arm in Section 7. The FER model associated with the perception module of the manipulator adapts to the user's facial expressions and personalizes the model when unsure of the new user's state of engagement.

The contributions of the paper lie in leveraging the feedback from high-accuracy ML models which have been built upon large datasets (in this case voice commands), to automatically label specialized human data (in this case FER data) for personalizing the latter to the behaviours of the human partner and the environment in which it has been deployed. The modes of communication used are incommensurable since paired data are challenging to find. Human feedback eliminates the possibility of mislabelling that is encountered when a third-party labels images of participants taken in a laboratory and acted out in a setting vastly different from the purported setting of deployment.

Emotions through facial expression recognition (FER) were selected for this study due to the additional complexities inherent to the adaptation task. While the framework would benefit all aspects of HRCom that use ML and need continuous learning (thus rendering it mode agnostic), affect data in particular would be far better recognized with correct labels generated from user feedback. Personalizing HRCom is, *ipso facto*, a promising effort towards safer systems. Additionally, the framework is an add-on created and deployed based on application-specific heuristics. There is no alteration of model architecture involved, thus making it model architecture agnostic.

# 4 Methodology

In an industrial setting, even for an all-human team, there are rarely a discernably large variety of emotions that are shown by humans; neutral expressions are the most common during task execution (Chiurco et al. 2022). That sets the application of FER system considered in this work apart from what may be required from a social human-robot interaction scenario. Having said that, the inclusion of negative emotions signifying the state of not being engaged is essential for safety.

As an implementation of the proposed framework that forms the contribution of the paper, a FER model is considered that has to detect whether the human is engaged in the task with their collaborative robot partner using RGB images. Inception v3 was selected to train the dataset of face images since it has demonstrated good performance in training for tasks involving learning human facial data (Chiurco et al. 2022; Tio 2019). The accuracy of ML models on FER is dependent on the angle of the face in images; models produce erroneous predictions if the entire face is not visible (Chiurco et al. 2022). This is a challenge due to the dynamic nature of working on a shop floor. Thus, it would be advantageous to train the model with a dynamic view of the face.

The facial expressions dataset used was a modified version of the Karolinska Directed Emotional Faces (KDEF) dataset (Kale et al. 2022). The KDEF dataset is a set of 4900 pictures of human facial expressions. The set of pictures contains 70 individuals displaying seven different emotional expressions. Each expression is viewed from five different angles (−90, −45, 0, +45, +90 degrees). The individuals are aged between 20 and 30 years and have no beards, mustaches, earrings, or eyeglasses. For the purpose of this study, the classes 'happy' and 'neutral' were considered as *Engaged* while 'anger', 'surprise', 'sad', 'afraid', and 'disgust' were considered as *Not Engaged*. The data was reorganized into train and test sets and balanced between the classes using data augmentation techniques. These augmentations included – lateral shifts so as to not overfit the model to expecting the human to be at the centre of the scene, zoom to mimic the varying distance of the human from the camera in a dynamic environment, and image rotations between −30 and 30 degrees to add some variance to the orientation of the head. In addition to that, since the lighting of all images was uniform, brightness shifts with delta between 0 and 0.8 were randomly added to the image pixels to allow for more variations in lighting conditions. The final modified dataset contained 7006 images. The advantage of using KDEF was the availability of the five viewpoints of the face that is expected to be representative of the dynamic nature of the work-cell wherein the human may not always be positioned at the centre of the camera that is capturing the images for FER.

As compared to the previously cited work (Faria et al. 2017) which also used KDEF but excluded images with the faces at 90 and −90 degrees due to the inadequate performance of facial landmark detection and geometrical feature detection algorithms, the model developed in this work using deep learning achieved an accuracy of 96.63% as compared to 85% of that work. A point to be clarified here is that while the models of (Faria et al. 2017) classified the dataset into seven classes, in this work, only two consolidated classes are used.

For the purposes of testing the framework for adaptation, a portion of user data was labelled by the user who participated in creating the dataset as test dataset, the images from which were not encountered by the models during training. The user (one of the authors) emoted in a way that seemed most natural to them instead of copying the posed expression from the KDEF dataset. Hence, the user data is believed to be more representative of their potential

behaviour in a collaborative scenario. As preprocessing, the images captured were cropped to contain some inches above the head, and some below the chin in a square image since rotations and small variations in the head and face were captured that needed some space laterally. These steps are heuristics and application-specific; the background may be blurred or the face detected before FER. Additionally, as mentioned in Section 2.1, segmentation of the face for better expression recognition may be applied. Conversely, if the collaboration task requires the model to be "aware" of the background of the human, then blurring may be unsafe. Thus, the specific preprocessing is up to the engineer, while the focus of this work is the framework that is generic.

The ML model trained on the KDEF dataset (*model0*) performed well on the KDEF test data (refer to Figure 1 (a)). But on testing (*model0*), on the two user datasets: 'Uncluttered User DS' (DS1) with the user's face and a clean background and 'Cluttered User DS' (DS2) with the same user but with noisy background, it was found to show accuracy scores of 76.44% and 81.19% (Figure 1 (b) and (c)). The high accuracy is due to the model bias towards *Not Engaged* class and due to the test datasets being unbalanced with the *Not Engaged* having more images than *Engaged*, thus skewing the accuracy measure and failing to recognize the other class images. The numbers of images in the test set in each were as follows.

- KDEF: *Engaged*: 700, *Not Engaged*: 698
- DS1: *Engaged*: 624, *Not Engaged*: 1647
- DS2: *Engaged*: 772, *Not Engaged*: 3332

The user images are out-of-distribution from the original KDEF dataset (Hendrycks and Gimpel 2016). From trivial human observation, the lighting conditions, the background are of course vastly different from the original. In a more non-trivial sense, the user data showcases expressions idiosyncratic of that person's behaviour and do not completely match with the posed ones in KDEF wherein every individual has expressed uniformly, something that is not representative of the real-world. Indeed, this strengthens the need to fine-tune the FER model based on the user data. Figure 2 contains a sample of images from KDEF, DS1 and DS2 grouped by labels.

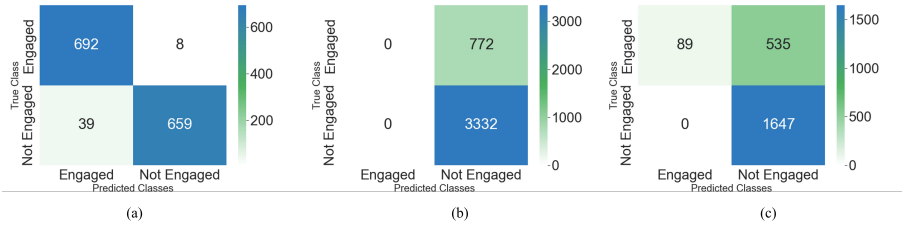


**Fig. 1** Performance of base model, *model0*, trained on KDEF only and tested on: (**a**) unseen KDEF data, (**b**) user data with cluttered background (DS2) and (**b**) user data with uncluttered background (DS1)

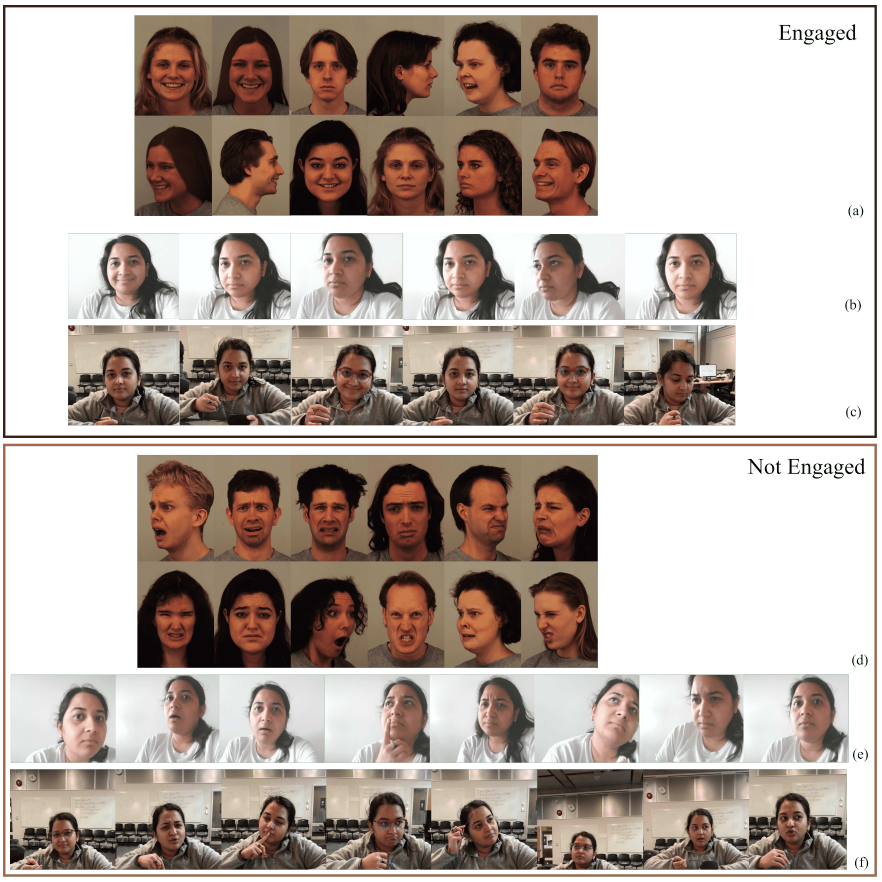


**Fig. 2** Images sampled from the three datasets considered: (**a**) images with *Engaged* label from KDEF, (**b**) images with *Engaged* label from user data with uncluttered background (DS1), (**c**) images with *Engaged* label from user data with cluttered background (DS2), (**d**) images with *Not Engaged* label from KDEF,(**e**) images with *Not Engaged* label DS1, (**f**) images with *Not Engaged* label DS2

The mode selected for generating labels of unannotated user data was voice command (VC). The ML model used for implementing VC classification was adapted from (Gajhede et al. 2016) with four convolution layers instead of three for better accuracy and a dropout layer to prevent overfitting. The dataset used was the speech commands dataset version 2 (Warden 2018). The two classes used were *yes* and *no*. The accuracy was 100% because the model was trained on a sufficiently large amount of data and the tests involved simple single-word commands of the user. Apart from voice commands, buttons on user interfaces or application-specific modes may also be used for generating feedback. In noisy factories, a graphical interface or hand gestures may be more suitable than voice commands, while voice commands may be more useful if the user's hands are otherwise occupied with the task or obscured.

On encountering shifted (user) data for the first time, verbal feedback was requested using VC classification through a direct question to the user, "Are you engaged in the task?". An answer of 'yes' generated label *Engaged* while 'no' generated *Not Engaged*, thus avoiding the labour-intensive task of hand labelling. *model0* was re-trained with the dataset of user images and generated labels to adapt or personalize to the user data without 'forgetting' the knowledge of the original dataset.

The first experiment was carried out to re-train the model with DS2 images from only one class along with images of KDEF of both classes. *model0* was trained using transfer learning with 20 images of *Engaged* class from DS2, 80 of *Engaged* from KDEF and 100 images of *Not Engaged* from KDEF to generate *model1*. All layers till the average pooling layer were frozen. The learning rate was 0.01 with the mini-batch size of 10, optimizer: stochastic gradient descent, the number of epochs of 2 for DS1 and 5 for DS2. *model1* was subsequently re-trained with 20 images from DS2 and 80 from KDEF of *Non Engaged* class and 100 from KDEF of *Engaged* class.

These two steps were repeated with new user data and the subsequent models were re-trained. On testing on the remaining DS2 dataset after each re-training, it was found that the newly generated model was completely biased towards the label from DS2 it had been trained on in that step (Figure 3). Odd numbered confusion matrices in (Figure 3) were generated on training with just label *Engaged* as described above while even-numbered ones were using the other label. Each model was built upon the previous one and was named to denote the cumulative process. In Figure 3, the numbering is as follows:

- 1,3,5,7: *Engaged*:20 DS2+ 80 KDEF, *Not Engaged*:100 KDEF generating *model1, model3, model5, model7* respectively
- 2,4,6,8: *Engaged*:100 KDEF, *Not Engaged*:20 DS2+ 80 KDEF generating *model2, model4, model6, model8* respectively

The new models, however, retained the knowledge of KDEF even after eight re-trainings with new data mixed in with the original data. It is posited that since the final fully connected layers contain only trainable parameters, and they are re-trained in every iteration, the generated models were unable to learn the features of images from both classes from the new dataset when data from only one class was passed for training. Since the models trained in the above-mentioned format were unable to learn both labels at once and showed bias during testing, the framework developed is based on iteratively re-training with small batches of new user data from both classes along with larger batches of old data. Re-Training in this document refers to transfer learning and not training from scratch

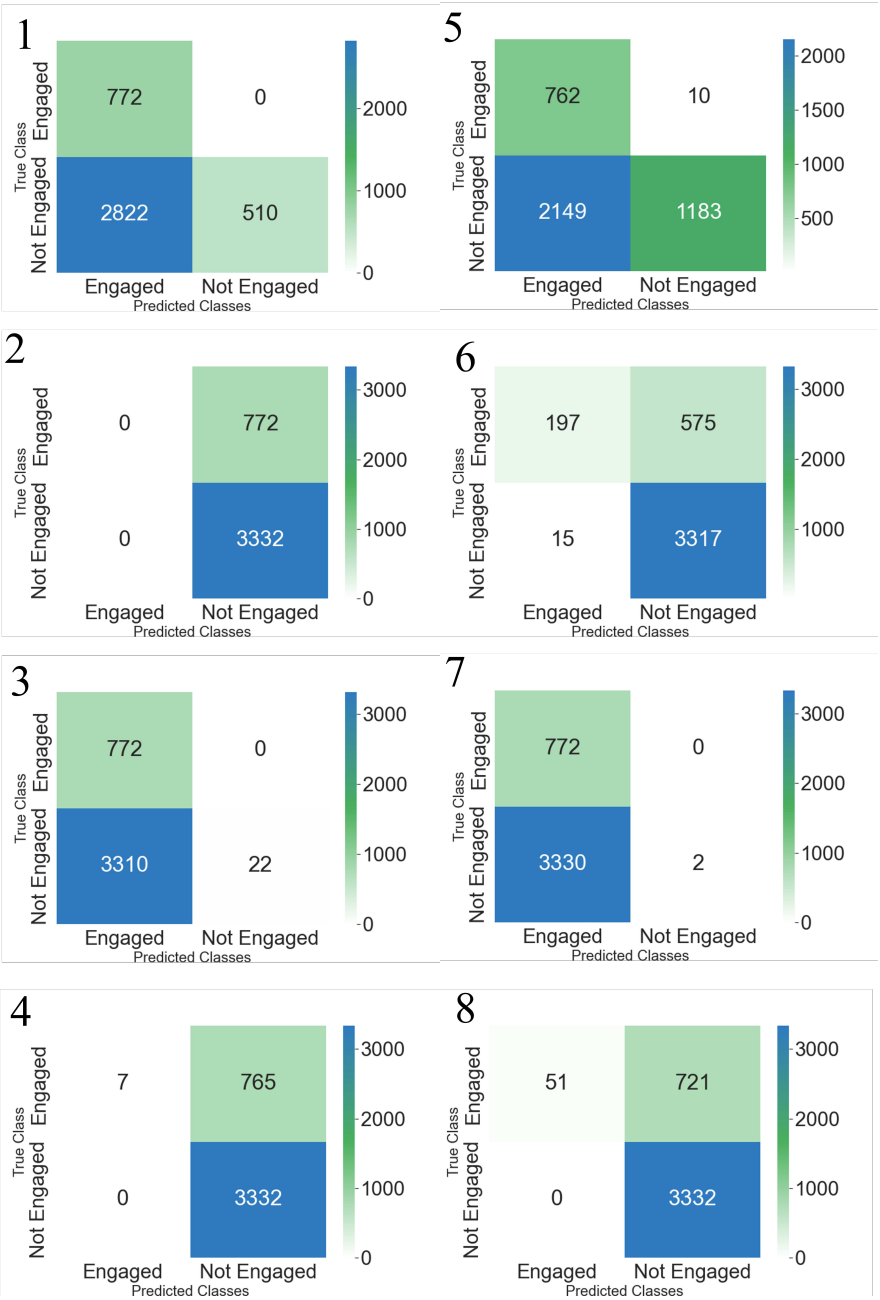


**Fig. 3** Confusion matrices generated from training on KDEF and subsequent re-training on DS2 with 20 images from only one label at a time from DS2, 80 of the same label from KDEF and 100 images from the other label from KDEF, and tested on unseen test images from DS2. Each model builds upon the previous one are named to denote the cumulative process. Note that models are biased towards whichever labels they were trained on with the user images.

Referring to the confusion matrix of Figure 1, it is to be noted that false positives (true class: *Not Engaged*, predicted class: *Engaged*) are undesirable since they can lead to safety issues, while false negatives (true class: *Engaged*, predicted class: *Not Engaged*) lead to lower efficiency. In order to balance both aspects of a collaborative scenario, F1-scores were tracked throughout the process.

## 4.1 Iterative Re-training PF-HRcom

Figure 4 presents a flowchart of PF-HRCom which consists of training with standard dataset(s) to generate the base model. The perception *scene* of the robot may be conceived as composed of two parts – the collaborating human and the setting or environment which includes the light(s), the sensor(s) for capturing communication, other interacting object(s) and artifacts such as walls, floor, ceiling, their colours and so on. A change in any component of the *scene* leads to a domain shift. Tracking the change of *scene* is a heuristic and practical process, e.g., if a new user is employed to work with the robot in the work-cell or the robot is placed in a different work-cell with the same human or both components change. The assumption is that the commands are common in all scenarios. This shifted data is collected in intervals of time that are specific to the application, stored in memory and then auto-labelled using feedback from a more robust communication mode. Subsequently, the model is re-trained with the small batch of "labelled" data and this process is repeated until the model learns the user data. This number of re-trainings needed will be specific to the application and domain of data.

The implementation presented in this paper consists of the case of the addition of a new user in the same environment and the change of the background environment of the work-cell along with the addition of a new user. The particular steps of the implemented PF-HRCom based on the flowchart are as follows:

1. Train an ML model with a standard, publicly-available dataset such that it performs well on ID data: base model (*model0* trained on KDEF).
2. If no component of the *scene* changes, deploy model.
3. If there is any change, to tackle the shift in data, collect the new user data and store in memory. This data is unlabelled since it gets generated in real-time by the user interacting with the robot.
4. Endeavouring to keep dataset size low, 20 images of the expression are then auto-labelled by the robot asking for verbal feedback from the user. A batch of 40 images, 20 of each class are stored in the memory that are then used to re-train the current model. The current model would be the base model in the first iteration.
5. The model is iteratively re-trained with the new batch of shifted user images. This was done to ensure effective learning of the model with a minimal amount of new data. The metrics are tracked and this process of

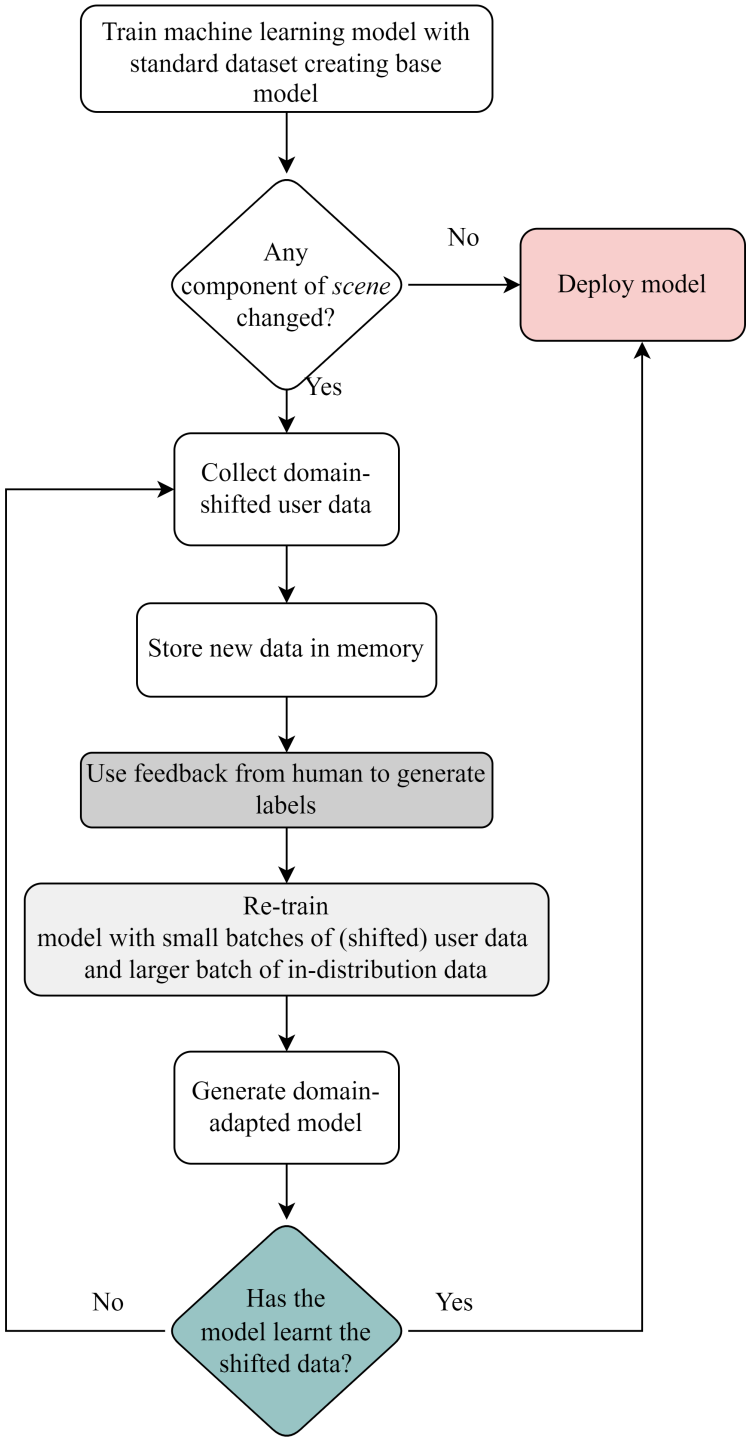


**Fig. 4** framework for leveraging multimodal input as feedback for tuning human-robot communication system using user-specific data

recording data, auto-labelling through verbal feedback using VC and re-training on the previous FER model is repeated till the threshold for a pre-set test metric is met or exceeded. In this implementation, testing of whether the newly generated model has learnt the features of the shifted data is tested using the F1 score, specifically, if the F1-score reaches above 0.8 or not.

The PF-HRCom was tested with both DS1 and DS2. DS2 was found to reach the requisite F1-score faster (and with a lesser amount of data) if trained on both DS1 and KDEF than if it was trained on just the base model. Since the batches of generated user data are so small, validation sets were not used, instead all the data were used for training. While the number of images from the user dataset was initially determined through trial and error of combinations for the purpose of implementation of the PF, ablation studies have been carried out in Section 6.

The other aspect to be noted is that this study used labelled DS1 and DS2 test sets to validate the framework in terms of if the newly generated model has learnt the features of the user data. F1 and accuracy scores were obtained on testing with those test sets. In real-world applications, such test sets would not be available and the process would have to be carried out in some other manner. Another way of understanding is that once a model has learnt the new data, it will not detect any shift in it. This brings us to the related aspect of the step for detection of change in *scene*. This is to be understood as the perception system being "unsure" of the engaged status of the human. An "engagement measure" using average of softmax probabilities have been used in the handover application in Section 7 as an automatic indication of whether the model has encountered shifted data. Once the model encounters such data, it pauses the robot operation and asks feedback for labelling and re-training.

# 5 Results and Discussion

For the purpose of testing the models and the framework, labelled datasets for both cluttered and uncluttered user data were created but the models were trained on images not in the test set. Table 1 contains the naming convention of the models generated through the experiments carried out. Table 2 and Figure 5 (a) contain the model evaluation metrics of iteratively labelling and re-training *model0* up to *model4* with uncluttered user data (DS1). Table 3 and Figure 5 (b) present the metrics of the re-training process with just cluttered one (DS2).

From Table 2, the row for *model0* shows the metrics of classification of DS1 with the base *model0* which has not been trained for DS1. It can be observed that *model0* is biased towards the *Not Engaged* class. This could be because the *Not Engaged* class contains images from five sub labels as mentioned earlier. Thus, it contains a higher variety of facial features. It seems reasonable that on encountering out-of-distribution faces and expressions (of the user), *model0* classifies them to *Not Engaged*. A similar behaviour is shown by *model0* tested

on DS2 as well (*model0* row of Table 3). The successive rows of Table 2 show a decrease in that bias after each model is generated from the earlier model along with an overall increase in F1-score and accuracy. This trend is followed in Table 3 as well for DS2 trained on the base model. This favourable behaviour is very dataset, domain, and application-specific and further bolsters the need for personalizing base models. The generation of new user data and re-training were stopped for DS1 and *model4* was used to test DS2. Table 4 presents the classification evaluation metrics on DS2 tested on *model0* and *model4* of DS1. It can be observed that although accuracy is similar for both models but higher F1-score and recall and overall lower number of false negatives render *model4* better than *model0*. This may be inferred to be because the same user was in both datasets, small as they may be.

To simulate the case in which the user has been included in the work-cell and the model has been adapted or personalized to that user and then the work-cell changes, *model4* from DS1 was used to train on DS2 with the same iterations as mentioned above. Through trial-and-error it was found that a combination of (20 DS2 + 40 DS1 + 40 KDEF) images for each class for iterative re-training yielded the best results (refer to Table 5). *model4* re-trained on DS2 with the given combination showcased 0.76 F1-score at an earlier stage than if it were to be trained just on KDEF. Without the KDEF data or without the samples from DS1, the re-trainings fared poorly, thus raising a need to keep the deployed FER model updated on the user data but slowly reducing the "standard" data as time of operation increases without becoming completely zero as ablation studies have demonstrated in Section 6. As more users get added to the work-cell the steps of PF-HRCom would need to be repeated for them for a more specific or personalized robot perception system.

Throughout the process, it was ensured through testing the generated models that the base knowledge i.e., of KDEF was not "forgotten" since that dataset contains much more variability in faces than the user data sets can. Thus, through the framework presented, the ensuing model contained less bias towards the *Not Engaged* class than the base model (refer to Table 5)) and showed higher classification accuracy and F1-score on sparse, out-of-distribution, unannotated user data with cluttered backgrounds.

As compared to the previously cited work (Faria et al. 2017) which also used KDEF the final personalized models trained with KDEF and user data developed in this work using PF-HRCom achieved higher accuracy scores than that of the models developed in the cited work which achieved accuracy score of around of 80% on user data outside of the dataset. To re-iterate, while the models of (Faria et al. 2017) classified the dataset into seven classes, in this work, only two consolidated classes are used.

**Table 1** Naming convention of models generated using the framework presented

| **Model** | **Transfer Learning using Model** | **Dataset** |
|---|---|---|
| *model0* | - | standard dataset (KDEF) |
| *model1 (for user dataset DSx)* | | |
| *x=1 for uncluttered dataset,* | *model0* | DSx batch_1 + KDEF |
| *x=2 for cluttered dataset* | | |
| *model2* | *model1* | DSx batch_2 + KDEF |
| *modeln* | *model(n-1)* | DSx batch_n + KDEF |

**Table 2** Iterative re-training of *model0* (trained on KDEF) on small batches of user data with uncluttered background (DS1) and testing on unseen DS1 images

| **Model** | **Precision** | **Recall** | **F1-Score** | **Accuracy(%)** | **False Positives** | **False Negatives** |
|---|---|---|---|---|---|---|
| *model0* | 1 | 0.14 | 0.25 | 76.44 | 0 | 535 |
| *model1* | 0.49 | 1.00 | 0.66 | 71.60 | 643 | 2 |
| *model2* | 0.71 | 0.97 | 0.82 | 88.42 | 246 | 17 |
| *model3* | 0.66 | 0.99 | 0.79 | 85.60 | 321 | 6 |
| *model4* | 0.88 | 0.95 | 0.91 | 95.07 | 83 | 29 |

**Table 3** Iterative re-training of *model0* (trained on KDEF) on small batches of user data with cluttered background (DS2) and testing on unseen DS2 images

| Model | Precision | Recall | F1-Score | Accuracy(%) | False Positives | False Negatives |
|---|---|---|---|---|---|---|
| *model0* | | 0 | | 81.19 | 0 | 772 |
| *model1* | 0.91 | 0.26 | 0.40 | 85.58 | 21 | 571 |
| *model2* | 0.90 | 0.58 | 0.71 | 90.86 | 51 | 324 |
| *model3* | 0.87 | 0.72 | 0.79 | 92.67 | 87 | 214 |
| *model4* | 0.81 | 0.82 | 0.81 | 92.96 | 153 | 136 |

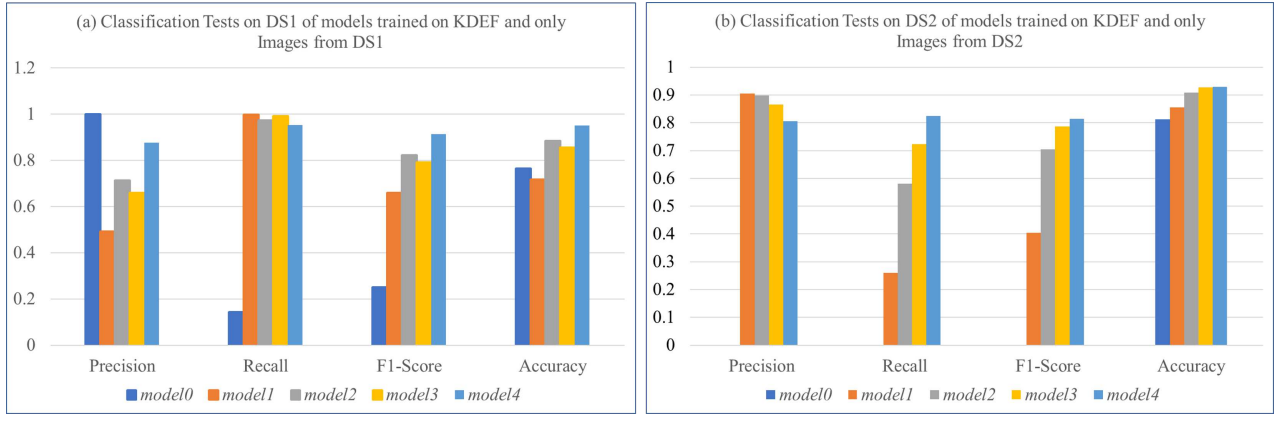


**Fig. 5** Plot of trends of metrics of (**a**)Classification Tests on DS1 of models trained on KDEF and only Images from DS1 and (**b**) Classification Tests on DS2 of models trained on KDEF and only Images from DS2

**Table 4** Testing of user data with cluttered background (DS2) on model trained with only KDEF (*model0*) and one trained on KDEF and iteraively on uncluttered user data (DS1) (*model4*)

| Model | Precision | Recall | F1-Score | Accuracy(%) | False Positives | False Negatives |
|---|---|---|---|---|---|---|
| *model0* | | 0 | | 81.19 | 0 | 772 |
| *model4* | 0.57 | 0.01 | 0.02 | 81.24 | 6 | 764 |

**Table 5** Iterative re-training of *model4* (trained on KDEF and uncluttered user data (DS1)) on small batches of user data with cluttered background (DS2) + KDEF + DS1 and testing on unseen DS2 images

| Model | Precision | Recall | F1-Score | Accuracy(%) | False Positives | False Negatives |
|---|---|---|---|---|---|---|
| *model4* | 0.57 | 0.01 | 0.02 | 81.24 | 6 | 764 |
| *model5* | 0.55 | 0.79 | 0.65 | 83.92 | 497 | 163 |
| *model6* | 0.67 | 0.89 | 0.77 | 89.84 | 331 | 86 |

# 6 Ablation Studies on Cluttered and Uncluttered User Datasets

While the implementation of the framework has been carried out with 20 images per label from the user dataset and 80 from KDEF, this section contains experiments of iterative re-training with the following combinations of images.

- Number of images per class from user dataset (either DS1 or DS2 depending on the experiment being carried out) = $m$

  $m \in \{10, 20, 30\}$ for DS1 and $m \in \{10, 20, 30, 40\}$ for DS2

- Number of images per class from KDEF (standard) dataset = $n$

  $n \in \{10n, 0 \leq n \leq 10 \forall n \in N\}$

  The three cases covered for this paper include the following:

1. Re-training model trained on KDEF with DS1 (uncluttered user dataset) only with the above-mentioned combinations of $m$ and $n$
2. Re-training model trained on KDEF with DS2 (cluttered user dataset) only with the above-mentioned combinations of $m$ and $n$
3. The conclusions from the ablation experiments carried out under 1 and 2 dictated the constitution of images per class for the experiments on re-training with DS2 on models already iteratively re-trained on DS1 and KDEF.

The rationale for selection of the number of images for user data stems from Paul Ekman's work that postulated that expressions last from 0.5-4 seconds (Ekman 2003). So, with our 20-fps camera, and in order to design a framework that would work with small batches of images, the number of images were selected starting from ($0.5s * 20fps = 10$ images) up to 40 for the more complex dataset (DS2). Combinations of user data and KDEF were carried out including case studies where no KDEF images were used for re-training and tuning.

## 6.1 Re-training model trained on KDEF with Uncluttered User Dataset (DS1) only

After the first iteration, for DS1, as number of images per class from KDEF ($n$) decreased, F1-scores also decreased for each case of $m$ (refer to Figure 6 (a), (d), and (g)). F1-scores with $m = 20$ were higher in most cases till $n = 40$. While the models retained learned features of KDEF in all combinations (as evidenced from high accuracy values on KDEF test set), the accuracy values on DS1 test set decreased with decrease in $n$ for all values of $m$ across all three iterations (Figure 6. (c), (f), and (i)). This may be because the variance in KDEF features allowed recognition of features in the *Not Engaged* class of DS1 which typically has the highest number of different expressions ('anger',

'surprise', 'sad', 'afraid', and 'disgust'). This may also explain the trend of lower F1-scores as $n$ was lowered across the experiments.

As $n$ decreased, accuracy values on the test set of KDEF also decreased (Figure 6 (d), (e), and (f)). This is an important consideration for re-training and tuning because excessive tuning of the model to a particular human would bias it and reduce the generalizability. This would negatively impact its practicality as more users get added to the work-cell or the environment changes.

For DS1, entries 12, 13, 23 and 25 in Table 6 with ($m$, $n$) combinations as (20, 100), (20, 90), (30, 100), (30, 80) reached the required F1-scores of $\geq 0.8$ with first iteration itself. These are highlighted in green while entries highlighted in blue are close ($0.77 \leq$ F1-scores $< 0.8$). In subsequent iterations, entries 37, 38 with just 20 total user images per class (10 per class in each iteration) and for entries 67, 70 and 71 with 30 total images (10+10+10) per class achieved the required metric. For higher accuracy values on DS1 and comparable or slightly better F1-scores, entries 78- 82 with $m = 20$ may also be considered which performed as well as using $m$=30. Indeed, the highest F1-scores were observed with ($m$,$n$): (20, 100) and (20, 80). Thus, values higher than 30 were not considered for $m$ in the case of DS1.

In general, the recommendation for similar datasets to DS1 (user data as close as possible to the standard data i.e., clean, solid colour background, no glasses or moustaches or other accessories and placement of the human in the image like the original) is to have higher number of images from the standard dataset ($n \in \{50, 60, 70,80, 90,100\}$) and $m = 20$ although higher values (30) and lower (10) would also be effective.

A point to be made here is that in the training setup, images were sampled randomly from a training set. This random sampling leads to stochasticity in test accuracy of the model trained on the sampled images. Thus, there are minimal variances in classification metrics for each run. While this random sampling has less of a discernible effect on DS1 experiments because of the simpler background and face, it shows a higher impact on DS2 because of the cluttered background and the presence or absence of user accessories such as glasses. Iterative re-training would lessen the impact on classification metrics as more data is added to train the model.

## 6.2 Re-training model trained on KDEF with Cluttered User Dataset (DS2) only

Accuracy scores on DS2 improved significantly after first re-training for all combinations of data (Table 7). The F1-scores were still below the target value of 0.8 hence, a second round of re-training was carried out using the models generated in the first round following the framework described in the previous sections. After first re-training, entries 127, 133-135, and 137-139 demonstrated F1-scores close to 0.7 highlighted in blue in Table 7. After second iteration, entries 155, 161, 162, 165, 171, 173-176, all combinations of $m = 40$ except entry 185 had F1-scores $\geq 0.7$. Out of these, entries 165, 176, and 187 have $n =$


**Fig. 6** Plot of trends of classification metrics of models generated from iterative re-training with uncluttered user dataset (DS1) with various combinations of DS1 and KDEF data

**Table 6** Classification metrics of models generated after iterative re-training with uncluttered user dataset (DS1) with various combinations of DS1 and KDEF over three iterations

| | Iteration 1 | | | | | Iteration 2 | | | | | Iteration 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry | No. of DS1 images /class (m) | No. of KDEF images/ class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS1 | Entry | No. of DS1 images/ class (m) | No. of KDEF images/ class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS1 | Entry | No. of DS1 images/ class (m) | No. of KDEF images /class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS1 |
| 1 | 10 | 100 | 0.95 | 0.83 | 0.75 | 34 | 10 | 100 | 0.97 | 0.88 | 0.74 | 67 | 10 | 100 | 0.97 | 0.92 | 0.85 |
| 2 | 10 | 90 | 0.97 | 0.88 | 0.78 | 35 | 10 | 90 | 0.97 | 0.88 | 0.74 | 68 | 10 | 90 | 0.97 | 0.89 | 0.78 |
| 3 | 10 | 80 | 0.96 | 0.87 | 0.78 | 36 | 10 | 80 | 0.97 | 0.87 | 0.74 | 69 | 10 | 80 | 0.96 | 0.88 | 0.73 |
| 4 | 10 | 70 | 0.97 | 0.86 | 0.70 | 37 | 10 | 70 | 0.97 | 0.91 | 0.83 | 70 | 10 | 70 | 0.97 | 0.89 | 0.81 |
| 5 | 10 | 60 | 0.97 | 0.85 | 0.70 | 38 | 10 | 60 | 0.96 | 0.90 | 0.81 | 71 | 10 | 60 | 0.97 | 0.92 | 0.84 |
| 6 | 10 | 50 | 0.95 | 0.78 | 0.71 | 39 | 10 | 50 | 0.97 | 0.85 | 0.71 | 72 | 10 | 50 | 0.97 | 0.87 | 0.79 |
| 7 | 10 | 40 | 0.95 | 0.83 | 0.75 | 40 | 10 | 40 | 0.97 | 0.86 | 0.77 | 73 | 10 | 40 | 0.97 | 0.86 | 0.77 |
| 8 | 10 | 30 | 0.92 | 0.65 | 0.61 | 41 | 10 | 30 | 0.96 | 0.87 | 0.77 | 74 | 10 | 30 | 0.96 | 0.86 | 0.77 |
| 9 | 10 | 20 | 0.95 | 0.82 | 0.75 | 42 | 10 | 20 | 0.95 | 0.85 | 0.76 | 75 | 10 | 20 | 0.96 | 0.85 | 0.77 |
| 10 | 10 | 10 | 0.94 | 0.78 | 0.70 | 43 | 10 | 10 | 0.94 | 0.85 | 0.75 | 76 | 10 | 10 | 0.95 | 0.86 | 0.76 |
| 11 | 10 | 0 | 0.91 | 0.73 | 0.62 | 44 | 10 | 0 | 0.91 | 0.71 | 0.61 | 77 | 10 | 0 | 0.91 | 0.73 | 0.61 |
| 12 | 20 | 100 | 0.97 | 0.90 | 0.82 | 45 | 20 | 100 | 0.97 | 0.90 | 0.83 | 78 | 20 | 100 | 0.96 | 0.94 | 0.88 |
| 13 | 20 | 90 | 0.96 | 0.89 | 0.82 | 46 | 20 | 90 | 0.97 | 0.90 | 0.83 | 79 | 20 | 90 | 0.97 | 0.90 | 0.83 |
| 14 | 20 | 80 | 0.96 | 0.80 | 0.73 | 47 | 20 | 80 | 0.96 | 0.82 | 0.74 | 80 | 20 | 80 | 0.95 | 0.93 | 0.87 |
| 15 | 20 | 70 | 0.96 | 0.85 | 0.78 | 48 | 20 | 70 | 0.96 | 0.84 | 0.76 | 81 | 20 | 70 | 0.94 | 0.91 | 0.82 |
| 16 | 20 | 60 | 0.96 | 0.85 | 0.77 | 49 | 20 | 60 | 0.96 | 0.88 | 0.80 | 82 | 20 | 60 | 0.96 | 0.89 | 0.81 |
| 17 | 20 | 50 | 0.96 | 0.86 | 0.78 | 50 | 20 | 50 | 0.96 | 0.85 | 0.77 | 83 | 20 | 50 | 0.96 | 0.88 | 0.77 |
| 18 | 20 | 40 | 0.95 | 0.83 | 0.76 | 51 | 20 | 40 | 0.96 | 0.85 | 0.77 | 84 | 20 | 40 | 0.96 | 0.85 | 0.77 |
| 19 | 20 | 30 | 0.95 | 0.80 | 0.73 | 52 | 20 | 30 | 0.95 | 0.84 | 0.75 | 85 | 20 | 30 | 0.95 | 0.86 | 0.77 |
| 20 | 20 | 20 | 0.94 | 0.77 | 0.70 | 53 | 20 | 20 | 0.95 | 0.88 | 0.78 | 86 | 20 | 20 | 0.94 | 0.84 | 0.75 |
| 21 | 20 | 10 | 0.93 | 0.71 | 0.65 | 54 | 20 | 10 | 0.94 | 0.84 | 0.73 | 87 | 20 | 10 | 0.94 | 0.82 | 0.72 |
| 22 | 20 | 0 | 0.90 | 0.72 | 0.62 | 55 | 20 | 0 | 0.90 | 0.72 | 0.62 | 88 | 20 | 0 | 0.91 | 0.77 | 0.62 |
| 23 | 30 | 100 | 0.96 | 0.88 | 0.80 | 56 | 30 | 100 | 0.96 | 0.88 | 0.80 | 89 | 30 | 100 | 0.96 | 0.90 | 0.83 |
| 24 | 30 | 90 | 0.96 | 0.86 | 0.79 | 57 | 30 | 90 | 0.95 | 0.91 | 0.82 | 90 | 30 | 90 | 0.96 | 0.91 | 0.83 |
| 25 | 30 | 80 | 0.95 | 0.88 | 0.80 | 58 | 30 | 80 | 0.94 | 0.92 | 0.83 | 91 | 30 | 80 | 0.96 | 0.91 | 0.84 |
| 26 | 30 | 70 | 0.95 | 0.78 | 0.71 | 59 | 30 | 70 | 0.96 | 0.84 | 0.76 | 92 | 30 | 70 | 0.95 | 0.92 | 0.85 |
| 27 | 30 | 60 | 0.95 | 0.76 | 0.70 | 60 | 30 | 60 | 0.94 | 0.89 | 0.81 | 93 | 30 | 60 | 0.96 | 0.90 | 0.83 |
| 28 | 30 | 50 | 0.95 | 0.83 | 0.75 | 61 | 30 | 50 | 0.95 | 0.80 | 0.71 | 94 | 30 | 50 | 0.95 | 0.91 | 0.84 |
| 29 | 30 | 40 | 0.95 | 0.84 | 0.76 | 62 | 30 | 40 | 0.95 | 0.86 | 0.78 | 95 | 30 | 40 | 0.95 | 0.86 | 0.78 |
| 30 | 30 | 30 | 0.94 | 0.79 | 0.72 | 63 | 30 | 30 | 0.94 | 0.89 | 0.76 | 96 | 30 | 30 | 0.95 | 0.82 | 0.74 |
| 31 | 30 | 20 | 0.94 | 0.78 | 0.70 | 64 | 30 | 20 | 0.94 | 0.82 | 0.74 | 97 | 30 | 20 | 0.94 | 0.83 | 0.74 |
| 32 | 30 | 10 | 0.94 | 0.74 | 0.66 | 65 | 30 | 10 | 0.93 | 0.78 | 0.69 | 98 | 30 | 10 | 0.93 | 0.79 | 0.70 |
| 33 | 30 | 0 | 0.90 | 0.73 | 0.62 | 66 | 30 | 0 | 0.91 | 0.75 | 0.63 | 99 | 30 | 0 | 0.91 | 0.83 | 0.74 |

0 i.e., no images of KDEF were used for re-training. These particular cases of $n = 0$ seemed to have lost significant knowledge of features learned from KDEF (Figure 7 (a)). As mentioned earlier, this indicates a decrease in variance that can be handled by the model and is evident from the lower accuracy values on tests on KDEF data. Hence, these entries are not considered as viable options for implementation of the framework.

The lack of increase of F1-scores for ($m$, $n$): (20, 10), (30, 40), (30, 70), and (30, 80) may possibly be explained by the stochasticity in the sampling process as mentioned earlier. In contrast, an increase in F1-scores can be observed for all other cases from iteration 1 to 2. Overall accuracy values on DS2 test images were higher for $m = 40$ while combinations with $m = 20$ and 30 were similar when accounting for stochasticity while the same for combinations with $m = 10$ were lowest. This follows since increase of training data will naturally lead to increase in accuracy if the sampled data is representative of the data to be learned. Even with a second and third re-training (entries 144-154 and 188-198), combinations with $m = 10$ showed slow increase in accuracy and F1-score values.

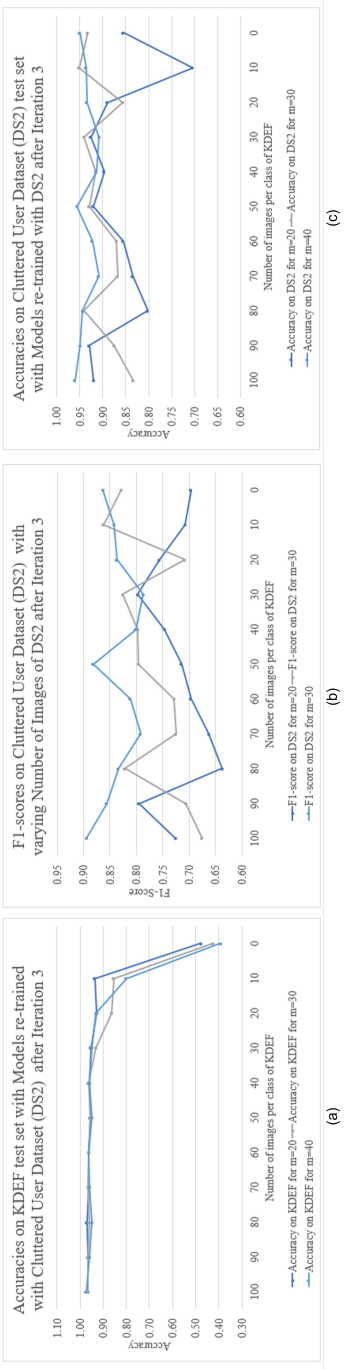


**Fig. 7** Plot of trends of classification metrics of models generated from iterative re-training with cluttered user dataset (DS2) with various combinations of DS2 and KDEF data

**Table 7** Classification metrics of models generated after iterative re-training with cluttered user dataset (DS2) with various combinations of DS2 and KDEF over three iterations

| | Iteration 1 | | | | | | Iteration 2 | | | | | | Iteration 3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry | No. of DS2 images/ class (m) | No. of KDEF images/ class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 | Entry | No. of DS2 images/ class (m) | No. of KDEF images/ class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 | Entry | No. of DS2 images/ class (m) | No. of KDEF images/ class (n) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 |
| 100 | 10 | 100 | 0.97 | 0.86 | 0.48 | 144 | 10 | 100 | 0.97 | 0.77 | 0.58 | 188 | 10 | 100 | 0.97 | 0.56 | 0.44 |
| 101 | 10 | 90 | 0.97 | 0.80 | 0.61 | 145 | 10 | 90 | 0.97 | 0.84 | 0.64 | 189 | 10 | 90 | 0.91 | 0.18 | 0.30 |
| 102 | 10 | 80 | 0.96 | 0.85 | 0.53 | 146 | 10 | 80 | 0.97 | 0.89 | 0.63 | 190 | 10 | 80 | 0.97 | 0.91 | 0.69 |
| 103 | 10 | 70 | 0.95 | 0.75 | 0.58 | 147 | 10 | 70 | 0.97 | 0.86 | 0.67 | 191 | 10 | 70 | 0.96 | 0.79 | 0.60 |
| 104 | 10 | 60 | 0.95 | 0.69 | 0.53 | 148 | 10 | 60 | 0.96 | 0.73 | 0.54 | 192 | 10 | 60 | 0.96 | 0.89 | 0.61 |
| 105 | 10 | 50 | 0.96 | 0.79 | 0.55 | 149 | 10 | 50 | 0.97 | 0.85 | 0.65 | 193 | 10 | 50 | 0.97 | 0.91 | 0.72 |
| 106 | 10 | 40 | 0.95 | 0.64 | 0.49 | 150 | 10 | 40 | 0.95 | 0.73 | 0.54 | 194 | 10 | 40 | 0.96 | 0.84 | 0.66 |
| 107 | 10 | 30 | 0.96 | 0.86 | 0.51 | 151 | 10 | 30 | 0.95 | 0.68 | 0.51 | 195 | 10 | 30 | 0.95 | 0.53 | 0.43 |
| 108 | 10 | 20 | 0.93 | 0.73 | 0.52 | 152 | 10 | 20 | 0.95 | 0.82 | 0.62 | 196 | 10 | 20 | 0.96 | 0.92 | 0.74 |
| 109 | 10 | 10 | 0.95 | 0.77 | 0.55 | 153 | 10 | 10 | 0.95 | 0.77 | 0.58 | 197 | 10 | 10 | 0.95 | 0.75 | 0.57 |
| 110 | 10 | 0 | 0.77 | 0.82 | 0.41 | 154 | 10 | 0 | 0.78 | 0.78 | 0.54 | 198 | 10 | 0 | 0.80 | 0.78 | 0.57 |
| 111 | 20 | 100 | 0.97 | 0.86 | 0.64 | 155 | 20 | 100 | 0.97 | 0.91 | 0.71 | 199 | 20 | 100 | 0.97 | 0.92 | 0.73 |
| 112 | 20 | 90 | 0.97 | 0.86 | 0.64 | 156 | 20 | 90 | 0.96 | 0.86 | 0.66 | 200 | 20 | 90 | 0.96 | 0.93 | 0.80 |
| 113 | 20 | 80 | 0.96 | 0.85 | 0.42 | 157 | 20 | 80 | 0.97 | 0.74 | 0.56 | 201 | 20 | 80 | 0.97 | 0.80 | 0.64 |
| 114 | 20 | 70 | 0.96 | 0.86 | 0.59 | 158 | 20 | 70 | 0.96 | 0.90 | 0.62 | 202 | 20 | 70 | 0.96 | 0.83 | 0.66 |
| 115 | 20 | 60 | 0.96 | 0.82 | 0.64 | 159 | 20 | 60 | 0.95 | 0.86 | 0.67 | 203 | 20 | 60 | 0.96 | 0.86 | 0.70 |
| 116 | 20 | 50 | 0.96 | 0.73 | 0.57 | 160 | 20 | 50 | 0.96 | 0.79 | 0.61 | 204 | 20 | 50 | 0.96 | 0.92 | 0.71 |
| 117 | 20 | 40 | 0.95 | 0.82 | 0.65 | 161 | 20 | 40 | 0.94 | 0.90 | 0.74 | 205 | 20 | 40 | 0.96 | 0.90 | 0.75 |
| 118 | 20 | 30 | 0.94 | 0.86 | 0.68 | 162 | 20 | 30 | 0.96 | 0.88 | 0.71 | 206 | 20 | 30 | 0.95 | 0.93 | 0.80 |
| 119 | 20 | 20 | 0.95 | 0.84 | 0.63 | 163 | 20 | 20 | 0.94 | 0.87 | 0.69 | 207 | 20 | 20 | 0.93 | 0.89 | 0.76 |
| 120 | 20 | 10 | 0.94 | 0.86 | 0.66 | 164 | 20 | 10 | 0.81 | 0.75 | 0.56 | 208 | 20 | 10 | 0.94 | 0.70 | 0.71 |
| 121 | 20 | 0 | 0.76 | 0.85 | 0.63 | 165 | 20 | 0 | 0.58 | 0.93 | 0.81 | 209 | 20 | 0 | 0.48 | 0.86 | 0.70 |
| 122 | 30 | 100 | 0.96 | 0.73 | 0.55 | 166 | 30 | 100 | 0.97 | 0.78 | 0.60 | 210 | 30 | 100 | 0.97 | 0.83 | 0.68 |
| 123 | 30 | 90 | 0.96 | 0.80 | 0.62 | 167 | 30 | 90 | 0.96 | 0.85 | 0.68 | 211 | 30 | 90 | 0.97 | 0.88 | 0.71 |
| 124 | 30 | 80 | 0.96 | 0.89 | 0.66 | 168 | 30 | 80 | 0.96 | 0.78 | 0.61 | 212 | 30 | 80 | 0.96 | 0.94 | 0.82 |
| 125 | 30 | 70 | 0.96 | 0.90 | 0.68 | 169 | 30 | 70 | 0.96 | 0.82 | 0.65 | 213 | 30 | 70 | 0.97 | 0.87 | 0.72 |
| 126 | 30 | 60 | 0.95 | 0.88 | 0.57 | 170 | 30 | 60 | 0.96 | 0.91 | 0.65 | 214 | 30 | 60 | 0.96 | 0.87 | 0.73 |
| 127 | 30 | 50 | 0.95 | 0.89 | 0.72 | 171 | 30 | 50 | 0.96 | 0.86 | 0.70 | 215 | 30 | 50 | 0.96 | 0.93 | 0.80 |
| 128 | 30 | 40 | 0.95 | 0.87 | 0.69 | 172 | 30 | 40 | 0.95 | 0.78 | 0.61 | 216 | 30 | 40 | 0.96 | 0.91 | 0.80 |
| 129 | 30 | 30 | 0.94 | 0.90 | 0.65 | 173 | 30 | 30 | 0.95 | 0.89 | 0.74 | 217 | 30 | 30 | 0.93 | 0.94 | 0.83 |
| 130 | 30 | 20 | 0.92 | 0.91 | 0.69 | 174 | 30 | 20 | 0.93 | 0.87 | 0.70 | 218 | 30 | 20 | 0.86 | 0.86 | 0.71 |
| 131 | 30 | 10 | 0.80 | 0.87 | 0.69 | 175 | 30 | 10 | 0.79 | 0.91 | 0.78 | 219 | 30 | 10 | 0.85 | 0.95 | 0.86 |
| 132 | 30 | 0 | 0.63 | 0.83 | 0.64 | 176 | 30 | 0 | 0.55 | 0.94 | 0.83 | 220 | 30 | 0 | 0.42 | 0.93 | 0.83 |
| 133 | 40 | 100 | 0.96 | 0.91 | 0.75 | 177 | 40 | 100 | 0.96 | 0.95 | 0.85 | 221 | 40 | 100 | 0.97 | 0.96 | 0.89 |
| 134 | 40 | 90 | 0.96 | 0.91 | 0.72 | 178 | 40 | 90 | 0.96 | 0.94 | 0.81 | 222 | 40 | 90 | 0.96 | 0.95 | 0.86 |
| 135 | 40 | 80 | 0.96 | 0.89 | 0.72 | 179 | 40 | 80 | 0.96 | 0.86 | 0.71 | 223 | 40 | 80 | 0.95 | 0.94 | 0.83 |
| 136 | 40 | 70 | 0.96 | 0.91 | 0.69 | 180 | 40 | 70 | 0.96 | 0.91 | 0.79 | 224 | 40 | 70 | 0.96 | 0.91 | 0.79 |
| 137 | 40 | 60 | 0.96 | 0.87 | 0.70 | 181 | 40 | 60 | 0.93 | 0.94 | 0.81 | 225 | 40 | 60 | 0.96 | 0.92 | 0.81 |
| 138 | 40 | 50 | 0.95 | 0.89 | 0.73 | 182 | 40 | 50 | 0.96 | 0.88 | 0.73 | 226 | 40 | 50 | 0.95 | 0.96 | 0.88 |
| 139 | 40 | 40 | 0.95 | 0.88 | 0.71 | 183 | 40 | 40 | 0.95 | 0.93 | 0.77 | 227 | 40 | 40 | 0.96 | 0.91 | 0.80 |
| 140 | 40 | 30 | 0.91 | 0.81 | 0.64 | 184 | 40 | 30 | 0.94 | 0.89 | 0.75 | 228 | 40 | 30 | 0.95 | 0.91 | 0.79 |
| 141 | 40 | 20 | 0.92 | 0.86 | 0.69 | 185 | 40 | 20 | 0.91 | 0.85 | 0.69 | 229 | 40 | 20 | 0.93 | 0.93 | 0.84 |
| 142 | 40 | 10 | 0.73 | 0.72 | 0.55 | 186 | 40 | 10 | 0.77 | 0.90 | 0.76 | 230 | 40 | 10 | 0.80 | 0.94 | 0.84 |
| 143 | 40 | 0 | 0.56 | 0.91 | 0.74 | 187 | 40 | 0 | 0.56 | 0.94 | 0.85 | 231 | 40 | 0 | 0.39 | 0.95 | 0.86 |

While combinations of $m = 40$ and $n \in \{$ 100,90,80,60,50, 40,0 $\}$ in the first iteration (entries 133-142) reached F1-scores $> 0.7$, after the second iteration, almost all combinations of $n$ with $m = 40$ reached the same metric with ($m$, $n$): (40, 100), (40, 90) and (40, 60) F1-scores $> 0.8$ which is the target for stopping tuning. With three iterative re-trainings, value of $n$ was independent when $m = 40$ for achieving the required criterion of F1-score $> 0.8$. The combination of (40, 0) is not viable since it lost the KDEF knowledge to a large extent. Similar results were obtained for batches with $m = 20$ (entries 155 and 161) which reached similarly high accuracy values with half the number of user images by the second iteration as compared to the combinations (40, 90) and (40, 60). Similarly for $m = 30$ with respect to entry 175. The third iteration yielded almost similar results for $m = 30$ and $m = 40$ (Table 7). With $m = 20$, the combinations of (20, 90) and (20, 30) were viable while most other combinations had F1-scores $> 0.7$ (highlighted in blue). From Figure 7, (c) the trends in the accuracy values on DS2 can be noted. The plot lines for $m = 30$ and 40 do not show as much variation as for $m = 20$, similar for (b) which are for F1-scores.

Overall, for a complex dataset such as DS2 which is quite different from the standard one used to train the original model, three or more iterations would be required with best results (high values of accuracy and F1-score) with small batches of $m = 30$ and independent of the number of KDEF images provided it is a non-zero value. Certain combinations with $m = 20$ may also work for this purpose, but they were subjected to stochasticity.

The $n = 0$ case displays very different behaviour for DS1 and DS2. The impact of the lack of standard data for re-training even over three iterations is quite minimal as compared to that for DS2 where accuracy on KDEF test set plummeted to 0.39 for $m = 40$ (entry 231, Table 7). This could perhaps indicate the similarity of DS1 to KDEF and could be an indicator of a means to expand datasets of this nature if the original dataset is not available for re-training of the model.

The attempt has been to keep the framework as close to a realistic scenario as possible. Hence, the collection and voice command feedback-based annotation of the user data has been undertaken in stages: of addition of a new user (within the same environment) and/or a change of environment. Change of environment with the same user could be considered as a superset of dataset of the scenario wherein the user is newly added to the old environment. The experiments in the presented paper have been demonstrated for this particular scenario. Thus, to present the methodology that uses as small batches of data as is feasible and realistic to generate within the application scenario, models for the new environment have been trained with only data generated in that environment and not a mixture of images from both scenarios. A final set of experiments have been carried out in subsection 6.3 wherein models already trained with DS1 were re-trained with DS2. As mentioned in Section 5, these models performed better than if only DS2 were used. The framework is flexible and if data is available, then the datasets from both scenarios may be merged as per the practitioner's discretion.

## 6.3 Model trained on DS1 used to iteratively re-train on DS2

From tests with combinations of DS1 (Section 6.1), it was found that $m$(DS1) = 20 and a high number of KDEF obtained best results, while $m$(DS2) = 30 and any non-zero number of KDEF images fared well (Section 6.2). DS1 is similar to KDEF while DS2 is more similar to DS1 (same human in the images) as compared to KDEF. Therefore, if models trained on DS1 (by transfer learning on the model trained with KDEF) were to be used to do transfer learning with DS2, a similar logic could be applied in formulating the combinations of images from the three datasets.

In case of DS2, $m \in \{ 10, 20, 30 \}$ were considered for tests and as per findings from Section 6.1, *n1* (images per class from DS1) $\in \{ 30, 40, 50 \}$ were used so as to keep high quantity of the more 'similar' dataset. Since results on DS2 were independent of the number of KDEF images used (Section 6.2), lower number of KDEF were used, *n2* (KDEF) $\in \{0,10,20,30,40 \}$. The combinations

arising from these values of *m*, *n1*, and *n2* were used to generate datasets for training for two iterations and then tested on test sets of DS2 and KDEF. Two iterations were run to test the hypothesis that models trained with KDEF, then DS1 would reach the requisite target F1-scores on DS2 through lesser re-training steps.

From Table 8 (entries 232-321), it can be observed that the combination of (*m*, *n1*, and *n2*): (30, 40, 20) in the first iteration and majority of combinations with *m*=30 after the second iteration reach F1-score $\geq$ 0.8 (entries 278-280, 284, 290-291, highlighted in green). The remaining (277, 282, 285, 286, 289) have F1-scores $\geq$ 0.77 (highlighted in blue) which are higher than with just DS2 and KDEF as covered in Section 6.2. The combination (20, 40, 40) also reaches the required target while (20, 40, 10) and (20, 40, 0) are close. (20, 40, 0) has no KDEF data during re-training, the effect of which is reflected in the lower accuracy of KDEF test set.

Outside of the combinations discussed above, $m = 30$, *n1* (DS2) $\in$ { 0, 10, 20 } and *n2* (KDEF) $\in$ { 10, 20 } were run as well, the results of which are captured in Table 8 (entries 322-333). Only entries 330 and 331 had F1-scores $\geq$ 0.77 after two iterations. From the table, it may be observed that combinations with $m = 30$ and *n1* $\geq$ 40 would be the ideal combination that reaches the required F1-score and accuracy on DS2 after lesser iterations than just using DS2 on KDEF and requires lesser number of images than certain cases discussed in Table 7 and Section 6.2 with *m*=40. The case of *n1*=0 does not seem to have a drastic impact on the classification metrics, but the long-term effects could be the subject of further study.

# 7 Application of PF-HRcom to a Manufacturing Task

The PF-HRCom framework was applied for a handover task from the robot to the human worker which is one of the common collaborative tasks in the industrial manufacturing site. The example process is designed to have a 6-DoF robot manipulator to hand over a tool to the human worker and a 20-fps stationary camera to capture the facial expression of the worker to ensure the engagement of the human.

The flow of the process is shown in Figure 8. When the process began, the robot received information from the manufacturing execution system about the current task such as product ID and determined the required tool. Then, the robot picked up the tool and moved it to the handover position following the pre-programmed sequence. During this sequence, the engagement of the human worker was simultaneously and continuously measured for each image frame with the current FER model. When the robot reached the handover position, FER is stopped and the system determines whether the worker is engaged or not throughout the robot's motion. If the human was *Engaged*, the robot completed the handover task by opening the gripper. If *Not Engaged*, the task was stopped and the robot put the tool back in its original position.

**Table 8** Classification metrics of models generated after iterative re-training with uncluttered user dataset (DS1) and then with cluttered user dataset (DS2)

| | Iteration 1 | | | | | | Iteration 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry | No. of DS2 images/ class (m) | No. of DS1 images/ class (n1) | No. of KDEF images/ class (n2) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 | Entry | No. of DS2 images/ class (m) | No. of DS1 images/ class (n1) | No. of KDEF images/ class (n2) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 |
| 232 | 30 | 50 | 40 | 0.94 | 0.91 | 0.72 | 277 | 30 | 50 | 40 | 0.95 | 0.91 | 0.78 |
| 233 | 30 | 50 | 30 | 0.94 | 0.86 | 0.70 | 278 | 30 | 50 | 30 | 0.95 | 0.93 | 0.81 |
| 234 | 30 | 50 | 20 | 0.93 | 0.88 | 0.74 | 279 | 30 | 50 | 20 | 0.92 | 0.93 | 0.80 |
| 235 | 30 | 50 | 10 | 0.90 | 0.91 | 0.73 | 280 | 30 | 50 | 10 | 0.93 | 0.93 | 0.82 |
| 236 | 30 | 50 | 0 | 0.86 | 0.92 | 0.73 | 281 | 30 | 50 | 0 | 0.76 | 0.85 | 0.69 |
| 237 | 30 | 40 | 40 | 0.95 | 0.92 | 0.78 | 282 | 30 | 40 | 40 | 0.95 | 0.91 | 0.78 |
| 238 | 30 | 40 | 30 | 0.94 | 0.86 | 0.68 | 283 | 30 | 40 | 30 | 0.94 | 0.92 | 0.76 |
| 239 | 30 | 40 | 20 | 0.90 | 0.92 | 0.81 | 284 | 30 | 40 | 20 | 0.94 | 0.95 | 0.85 |
| 240 | 30 | 40 | 10 | 0.92 | 0.88 | 0.73 | 285 | 30 | 40 | 10 | 0.85 | 0.90 | 0.77 |
| 241 | 30 | 40 | 0 | 0.90 | 0.86 | 0.70 | 286 | 30 | 40 | 0 | 0.77 | 0.93 | 0.78 |
| 242 | 30 | 30 | 40 | 0.95 | 0.83 | 0.66 | 287 | 30 | 30 | 40 | 0.94 | 0.89 | 0.65 |
| 243 | 30 | 30 | 30 | 0.94 | 0.90 | 0.75 | 288 | 30 | 30 | 30 | 0.95 | 0.86 | 0.71 |
| 244 | 30 | 30 | 20 | 0.91 | 0.87 | 0.71 | 289 | 30 | 30 | 20 | 0.93 | 0.92 | 0.78 |
| 245 | 30 | 30 | 10 | 0.92 | 0.88 | 0.72 | 290 | 30 | 30 | 10 | 0.88 | 0.93 | 0.81 |
| 246 | 30 | 30 | 0 | 0.90 | 0.91 | 0.70 | 291 | 30 | 30 | 0 | 0.77 | 0.93 | 0.83 |
| 247 | 20 | 50 | 40 | 0.95 | 0.89 | 0.72 | 292 | 20 | 50 | 40 | 0.95 | 0.90 | 0.73 |
| 248 | 20 | 50 | 30 | 0.93 | 0.86 | 0.69 | 293 | 20 | 50 | 30 | 0.94 | 0.91 | 0.71 |
| 249 | 20 | 50 | 20 | 0.94 | 0.85 | 0.69 | 294 | 20 | 50 | 20 | 0.84 | 0.92 | 0.72 |
| 250 | 20 | 50 | 10 | 0.90 | 0.91 | 0.70 | 295 | 20 | 50 | 10 | 0.93 | 0.86 | 0.72 |
| 251 | 20 | 50 | 0 | 0.91 | 0.91 | 0.71 | 296 | 20 | 50 | 0 | 0.78 | 0.83 | 0.65 |
| 252 | 20 | 40 | 40 | 0.95 | 0.90 | 0.71 | 297 | 20 | 40 | 40 | 0.94 | 0.92 | 0.80 |
| 253 | 20 | 40 | 30 | 0.94 | 0.88 | 0.71 | 298 | 20 | 40 | 30 | 0.94 | 0.89 | 0.74 |
| 254 | 20 | 40 | 20 | 0.94 | 0.89 | 0.71 | 299 | 20 | 40 | 20 | 0.92 | 0.91 | 0.73 |
| 255 | 20 | 40 | 10 | 0.92 | 0.88 | 0.67 | 300 | 20 | 40 | 10 | 0.93 | 0.91 | 0.79 |
| 256 | 20 | 40 | 0 | 0.86 | 0.84 | 0.68 | 301 | 20 | 40 | 0 | 0.85 | 0.93 | 0.78 |
| 257 | 20 | 30 | 40 | 0.95 | 0.89 | 0.68 | 302 | 20 | 30 | 40 | 0.95 | 0.81 | 0.64 |
| 258 | 20 | 30 | 30 | 0.93 | 0.67 | 0.52 | 303 | 20 | 30 | 30 | 0.94 | 0.84 | 0.68 |
| 259 | 20 | 30 | 20 | 0.92 | 0.86 | 0.64 | 304 | 20 | 30 | 20 | 0.94 | 0.91 | 0.68 |
| 260 | 20 | 30 | 10 | 0.92 | 0.84 | 0.67 | 305 | 20 | 30 | 10 | 0.92 | 0.91 | 0.75 |
| 261 | 20 | 30 | 0 | 0.84 | 0.84 | 0.65 | 306 | 20 | 30 | 0 | 0.75 | 0.92 | 0.78 |
| 262 | 10 | 50 | 40 | 0.94 | 0.86 | 0.65 | 307 | 10 | 50 | 40 | 0.95 | 0.88 | 0.61 |
| 263 | 10 | 50 | 30 | 0.94 | 0.86 | 0.64 | 308 | 10 | 50 | 30 | 0.94 | 0.85 | 0.68 |
| 264 | 10 | 50 | 20 | 0.93 | 0.88 | 0.64 | 309 | 10 | 50 | 20 | 0.94 | 0.77 | 0.59 |
| 265 | 10 | 50 | 10 | 0.92 | 0.89 | 0.58 | 310 | 10 | 50 | 10 | 0.93 | 0.75 | 0.58 |
| 266 | 10 | 50 | 0 | 0.92 | 0.89 | 0.66 | 311 | 10 | 50 | 0 | 0.81 | 0.69 | 0.53 |
| 267 | 10 | 40 | 40 | 0.95 | 0.87 | 0.57 | 312 | 10 | 40 | 40 | 0.95 | 0.90 | 0.71 |
| 268 | 10 | 40 | 30 | 0.95 | 0.82 | 0.63 | 313 | 10 | 40 | 30 | 0.95 | 0.89 | 0.68 |
| 269 | 10 | 40 | 20 | 0.94 | 0.87 | 0.63 | 314 | 10 | 40 | 20 | 0.85 | 0.90 | 0.61 |
| 270 | 10 | 40 | 10 | 0.87 | 0.80 | 0.61 | 315 | 10 | 40 | 10 | 0.93 | 0.82 | 0.64 |
| 271 | 10 | 40 | 0 | 0.92 | 0.89 | 0.63 | 316 | 10 | 40 | 0 | 0.85 | 0.56 | 0.44 |
| 272 | 10 | 30 | 40 | 0.94 | 0.81 | 0.61 | 317 | 10 | 30 | 40 | 0.96 | 0.78 | 0.59 |
| 273 | 10 | 30 | 30 | 0.95 | 0.88 | 0.60 | 318 | 10 | 30 | 30 | 0.95 | 0.82 | 0.64 |
| 274 | 10 | 30 | 20 | 0.94 | 0.82 | 0.63 | 319 | 10 | 30 | 20 | 0.94 | 0.86 | 0.71 |
| 275 | 10 | 30 | 10 | 0.94 | 0.86 | 0.59 | 320 | 10 | 30 | 10 | 0.86 | 0.88 | 0.70 |
| 276 | 10 | 30 | 0 | 0.82 | 0.89 | 0.69 | 321 | 10 | 30 | 0 | 0.79 | 0.88 | 0.72 |

| | Iteration 1 | | | | | | Iteration 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entry | No. of DS2 images/ class (m) | No. of DS1 images/ class (n1) | No. of KDEF images/ class (n2) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 | Entry | No. of DS2 images/ class (m) | No. of DS1 images/ class (n1) | No. of KDEF images/ class (n2) | Accuracy on KDEF | Accuracy on DS2 | F1-score on DS2 |
| 322 | 30 | 20 | 20 | 0.95 | 0.91 | 0.72 | 328 | 30 | 20 | 20 | 0.95 | 0.90 | 0.76 |
| 323 | 30 | 10 | 20 | 0.93 | 0.89 | 0.69 | 329 | 30 | 10 | 20 | 0.95 | 0.91 | 0.75 |
| 324 | 30 | 0 | 20 | 0.90 | 0.92 | 0.72 | 330 | 30 | 0 | 20 | 0.95 | 0.90 | 0.77 |
| 325 | 30 | 20 | 10 | 0.91 | 0.85 | 0.69 | 331 | 30 | 20 | 10 | 0.90 | 0.93 | 0.78 |
| 326 | 30 | 10 | 10 | 0.91 | 0.89 | 0.70 | 332 | 30 | 10 | 10 | 0.81 | 0.88 | 0.72 |
| 327 | 30 | 0 | 10 | 0.84 | 0.91 | 0.71 | 333 | 30 | 0 | 10 | 0.79 | 0.86 | 0.71 |

In case the perception system encountered shifted data due to any change in *scene*, the system would be "unsure" of the human being either *Engaged* or *Not Engaged*, the operation paused. The system then asked for user feedback and the PF-HRCom framework was executed to update the FER model using the newly achieved information.

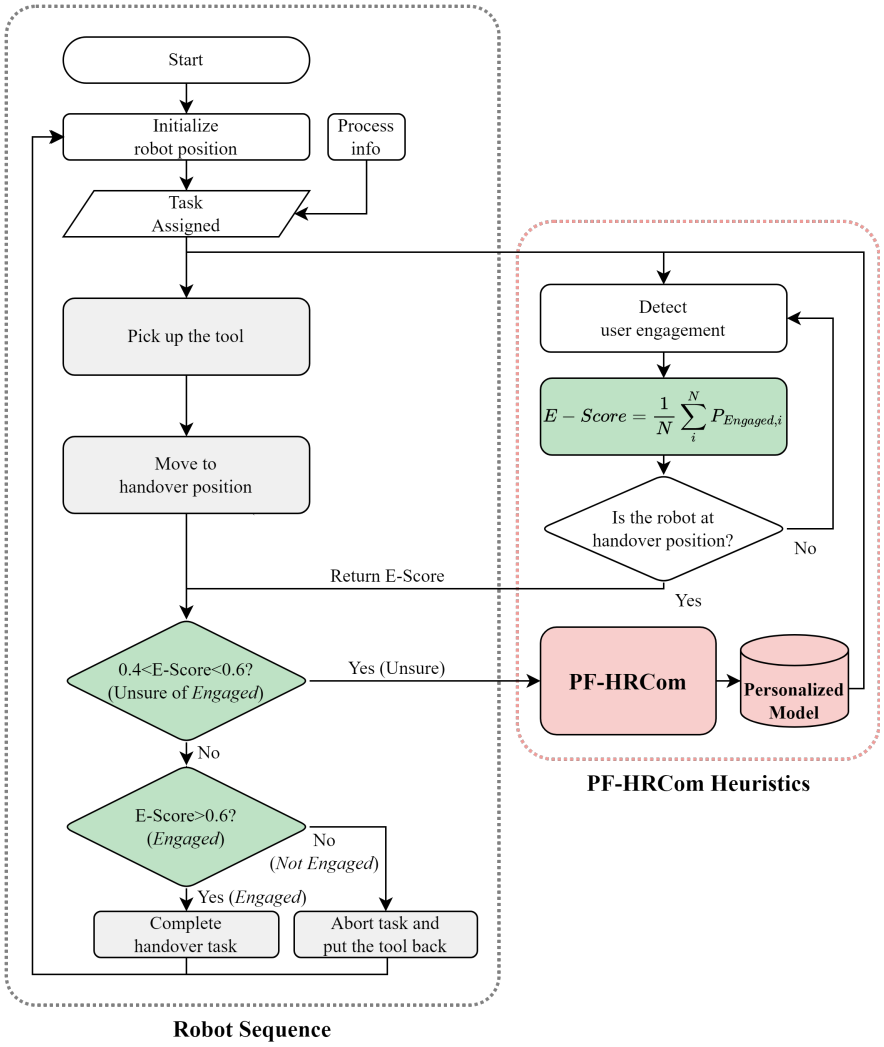


**Fig. 8** The flowchart of the example collaborative task for robot-human tool handover scenario using PF-HRCom framework.

In this application, the softmax probabilities for *Engaged* class $P_{Engaged}$ across the image frames are averaged to determine the engagement score(E-Score):

$$E - Score = \frac{1}{N} \sum_{i}^{N} P_{Engaged,i}$$

where N is the number of image frames. The threshold of the E-Score for *Engaged* is set to $E - Score \geq 0.6$, whereas the threshold for *Not Engaged* is set to $E - Score < 0.4$. If $0.4 < E - Score \leq 0.6$, the FER is considered unsure of facial expression because of *scene* changes out of the distribution from the previous FER model(e.g. a new worker is captured). These thresholds are the heuristic values and can be adjusted by the system designer. Also, other metrics for E-Score other than the average of softmax probabilities can be applied to a further extent in this scheme to identify the unsure cases in a safer manner.

The application consisted of two participants (authors of this work) who will now be referred to as User1 and User2 (workers in the industrial task) in this industrial scenario. User2 was not a part of the FER analysis presented in Sections 4 and 5. The initial model (*model1*) was trained with KDEF (standard) dataset and data of User1. The data for User1 was captured in the same environment on-site and with the robot in the work-cell. During the application of PF-HRCom, the following cases arose:

***Case1*** User1 was engaged in the handover task, verified by E-Score. *model1* which was trained with User1 data recognized the *Engaged* label and the handover task was successfully completed.
***Case2*** User1 was not engaged in the handover task, verified by E-Score. *model1* which was trained with User1 data recognized the *Not Engaged* label and the handover task was stopped for the safety of the worker.
***Case3*** User2 was engaged in the handover task. But *model1* which was not trained with User2 data was unable to come to a decision on the *Engaged* status label. For safety, the task was paused and a GUI shown on the screen asked the worker, "Are you engaged"? The feedback of "yes" or "no" would be used to label the data for re-training. *model2* was generated by training on User2 data in addition to User1 and KDEF, thus personalizing to the second user as well.

After generating a more personalized model: *model2*, the setup was run again and the following cases arose:

***Case4*** User1 was engaged in the handover task, verified by E-Score. *model2* recognized the *Engaged* label and the handover task was completed.
***Case5*** User1 was not engaged in the handover task, verified by E-Score. *model2* recognized the *Not Engaged* label and the handover task was stopped for the safety of the worker.
***Case6*** User2 was engaged in the handover task, verified by E-Score. *model2* was able to recognize the *Engaged* label and the handover task was successfully

completed, thus showcasing the adapted behaviour of the robot perception system to the new worker.
***Case7*** User2 was not engaged in the handover task, verified by E-Score. *model2* recognized the *Not Engaged* label and the handover task was stopped for the safety of the worker.

The case of User2 in a not engaged state was not carried out since as mentioned earlier, the FER model is biased towards the *Not Engaged* label and was thus considered redundant. A video of the application can be accessed from: https://drive.google.com/file/d/1P4sad2GW-OeAqt7bKdkoAJ9WlfIACCcF/view?usp=share_link. Figure 9 contains screenshots of the video organized under the seven cases.

## 8 Conclusions and Further Work

This paper presents a novel framework – PF-HRCom for the development of a personalized robot perception system. It has been validated upon facial expression recognition (FER) with sparse data from the human partner. An in-commensurable mode of communication, the voice command from the user was leveraged to annotate the user data and iteratively re-train the FER model without having to hand-label the datasets. PF-HRCom is further applied to a human-robot handover task that takes facial expressions as indications of the focus of the human in the task. The process highlighted the need to analyze application and domain-specific data and the need to customize machine learning models to the user and the requirement. The high classification metrics of the trained model tested on user data were better than the model trained with the publicly available "standard" dataset. This was demonstrated on datasets with both clean as well as noisy, cluttered backgrounds.

The framework has been demonstrated and presented using voice commands to provide feedback, but it is mode agnostic that is to say, physiological signals or body language recognition models may also be used for feedback. Indeed, the use of the said modalities may generate higher accuracy and lesser time of detection of the human's internal states. Also, instead of RGB images, videos may also be used for recognition of 'engagement' of the human in the task. In addition to that, hand gestures or body poses specific to certain applications that also suffer from the absence of large, labelled datasets may be generated and models trained using this framework, the applicability of which is not just limited to affect data.

While E-Score was an easy-to-use metric that was used to test the model performance in the application section, it is to be noted that a similar approach may not be suitable for a multiclass classification model. This is because neural network classifiers tend to classify OODs with high confidence resulting from the use of SoftMax function. This is because the probabilities are computed using exponential functions which lead to a large increase in output for a minor increase in input (Gal and Ghahramani 2016). Thus, more involved confidence or domain shift quantification measures may be required for such

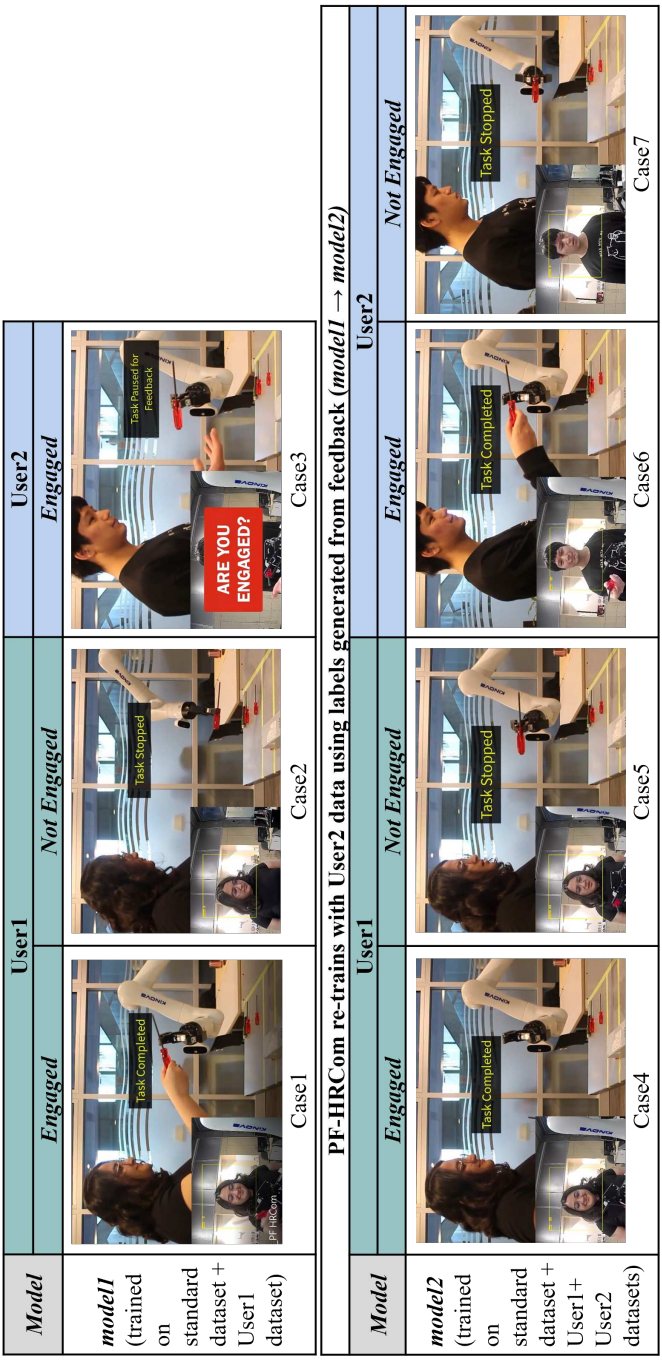


**Fig. 9** Application of Personalization through Feedback enabled Human-Robot Communication (PF-HRCom) framework to a manufacturing task

cases. Statistically driven unsupervised approaches to gauge model learning may be a lucrative future work.

The complexity of human emotions may be tackled to provide greater nuance to HRCom. Intensity or arousal of emotions (Citron et al. 2014) can be used to increase the specificity of the model to the human by means of the framework. With personalized baselines of intensity of emotions for each user, any change in emotion during the task that can distract the human can be better recognised by the system. Furthermore, checkpoints may be added while implementing PF-HRCom to take feedback and re-train if negative interactions are being recognised. This would reduce the downtime of the robot and enable better personalization and higher efficiency. Implementing such scenarios is an interesting future work from the perspective of affective computing.

The case of misclassification by the model may be a safety issue. This has not been tackled in this work since it would be a rare occurrence because the FER models were biased towards *Not Engaged* classes, thus would stop operation even due to misclassification, and due to the high accuracy of said models. Future work in other modes of communication could be more susceptible to misclassification. Fusion of multiple modes of communication may be used instead of just one model to tackle this issue.

Finally, human-robot collaboration and more generally, human-machine interaction systems designed to carry out complex manipulation tasks must also involve a robot perception system that can adapt itself to the human. The latter will support communication and lead to higher safety and more efficiency in communication.

## Declarations

- Funding: Research supported by UBC Office of the Vice-President, Research and Innovation in the form of seed funding to establish research on digitalization of manufacturing and Mitacs Globalink Research Internship award, 2020.
- Conflict of interest: The authors have no relevant financial or non-financial interests to disclose.
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: All authors have read and agreed to the published version of the manuscript.
- Availability of data and materials: The data that support the findings of this study are available from the corresponding author upon request.
- Authors' contributions:
  Debasmita Mukherjee: Development of Methodology presented, Formal analysis, Data curation, Investigation, Software development of implemented framework and FER model, Visualization, FER application with robot, Writing-original draft;
  Jayden Hong: Robot implementation, Writing of Section 7, Visualization;

Haripriya Vats: Data curation, Software development of FER model;
Sooyeon Bae: Software development of VC classification model;
Homayoun Najjaran: Funding acquisition, Project administration, Supervision, Writing- review & editing

## References

Affectiva (2018). Building the ultimate in-cabin experience with renovo and affectiva.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68. PMID: 31313636.

Caleb-Solly, P., Dogramadzi, S., Huijnen, C. A., and van den Heuvel, H. (2018). Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance. *The Information Society*, 34(3):153–165.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets.

Chen, L., Zhou, M., Su, W., Wu, M., She, J., and Hirota, K. (2018). Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61.

Chi, S., Tian, Y., Wang, F., Zhou, T., Jin, S., and Li, J. (2022). A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. *Artificial Intelligence in Medicine*, 125:102256.

Chiurco, A., Frangella, J., Longo, F., Nicoletti, L., Padovano, A., Solina, V., Mirabelli, G., and Citraro, C. (2022). Real-time detection of worker's emotions for advanced human-robot interaction during collaborative tasks in smart factories. *Procedia Computer Science*, 200:1875–1884. 3rd International Conference on Industry 4.0 and Smart Manufacturing.

Churamani, N., Anton, P., Brügger, M., Fließwasser, E., Hummel, T., Mayer, J., Mustafa, W., Ng, H. G., Nguyen, T. L. C., Nguyen, Q., Soll, M., Springenberg, S., Griffiths, S., Heinrich, S., Navarro-Guerrero, N., Strahl, E., Twiefel, J., Weber, C., and Wermter, S. (2017). The impact of personalisation on human-robot interaction in learning scenarios. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, page 171–180, New York, NY, USA. Association for Computing Machinery.

Citron, F. M., Gray, M. A., Critchley, H. D., Weekes, B. S., and Ferstl, E. C. (2014). Emotional valence and arousal affect reading in an interactive way: neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, 56:79–89.

Di Napoli, C., Valentino, M., Sabatucci, L., and Cossentino, M. (2018). Adaptive workflows of home-care services. In *2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 3–8.

Drawdy, C. C. and Yanik, P. M. (2015). Gaze estimation technique for directing assistive robotics. *Procedia Manufacturing*, 3:837–844. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* Emotions revealed: Recognizing faces and feelings to improve communication and emotional life. Times Books/Henry Holt and Co, New York, NY, US. Pages: xvii, 267.

Ekman, P. and Friesen, W. V. (2003). *Unmasking the face.* Malor Books, Cambridge, Mass.

Faria, D. R., Vieira, M., Faria, F. C., and Premebida, C. (2017). Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 805–810. IEEE.

Gajhede, N., Beck, O., and Purwins, H. (2016). Convolutional neural networks with batch normalization for classifying hi-hat, snare, and bass percussion sound samples. In *Proceedings of the Audio Mostly 2016*, AM '16, page 111–115, New York, NY, USA. Association for Computing Machinery.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning

contests.

Gupta, S., Hoffman, J., and Malik, J. (2015). Cross modal distillation for supervision transfer.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv*.

Hsu, S.-C., Huang, H.-H., and Huang, C.-L. (2017). Facial expression recognition for human-robot interaction. In *2017 First IEEE International Conference on Robotic Computing (IRC)*, pages 1–7.

Kale, Y. V., Shetty, A. U., Patil, Y. A., Patil, R. A., and Medhane, D. V. (2022). Object detection and face recognition using yolo and inception model. In Woungang, I., Dhurandher, S. K., Pattanaik, K. K., Verma, A., and Verma, P., editors, *Advanced Network Technologies and Intelligent Computing*, pages 274–287, Cham. Springer International Publishing.

Kardos, C., Kemény, Z., Kovács, A., Pataki, B. E., and Váncza, J. (2018). Context-dependent multimodal communication in human-robot collaboration. *Procedia CIRP*, 72:15–20.

Khan, O., Badhiwala, J. H., Grasso, G., and Fehlings, M. G. (2020). Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care. *World Neurosurgery*, 140:512–518.

Kim, D. Y. and Wallraven, C. (2021). Label quality in affectnet: results of crowd-based re-annotation.

Kim, J.-B. and Park, J.-S. (2016). Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. *Engineering applications of artificial intelligence*, 52:126–134.

Kothandaraman, D., Nambiar, A., and Mittal, A. (2020). Domain adaptive knowledge distillation for driving scene semantic segmentation.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

Kumagai, K., Lin, D., Meng, L., Blidaru, A., Beesley, P., Kulić, D., and Mizuuchi, I. (2018). Towards individualized affective human-machine interaction. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 678–685.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft coco: Common objects in context.

Liu, H., Fang, T., Zhou, T., and Wang, L. (2018). Towards robust human-robot collaborative manufacturing: Multimodal fusion. *IEEE Access*, 6:74762–74771.

Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M., and Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676.

Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101.

Maroto-Gómez, M., Marqués-Villaroya, S., Castillo, J. C., Álvaro Castro-González, and Malfaz, M. (2023). Active learning based on computer vision and human–robot interaction for the user profiling and behavior personalization of an autonomous social robot. *Engineering Applications of Artificial Intelligence*, 117:105631.

Maurtua, I., Fernandez, I., Kildal, J., Susperregi, L., Tellaeche, A., and Ibarguren, A. (2016). Enhancing safe human-robot collaboration through natural multimodal communication. In *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8.

Maurtua, I., Fernández, I., Tellaeche, A., Kildal, J., Susperregi, L., Ibarguren, A., and Sierra, B. (2017). Natural multimodal communication for human–robot collaboration. *International Journal of Advanced Robotic Systems*, 14(4):1729881417716043.

Mohammed, S. N. and Hassan, A. K. A. (2020). A survey on emotion recognition for human robot interaction. *Journal of computing and information technology*, 28(2):125–146.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Mozilla (2022). Mozilla common voice, https://voice.mozilla.org/en.

Mukherjee, D., Gupta, K., Chang, L. H., and Najjaran, H. (2022a). A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing*, 73:102231.

Mukherjee, D., Gupta, K., and Najjaran, H. (2022b). An ai-powered hierarchical communication framework for robust human-robot collaboration in industrial settings. In *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN),accepted, in press*, pages 1–6.

Mukherjee, D., Gupta, K., and Najjaran, H. (2022c). A critical analysis of industrial human-robot communication and its quest for naturalness through the lens of complexity theory. *Frontiers in Robotics and AI*, 9.

Nuzzi, C., Pasinetti, S., Pagani, R., Ghidini, S., Beschi, M., Coffetti, G., and Sansoni, G. (2021). Meguru: a gesture-based robot program builder for meta-collaborative workstations. *Robotics and Computer-Integrated Manufacturing*, 68:102085.

Rautiainen, S., Pantano, M., Traganos, K., Ahmadi, S., Saenz, J., Mohammed, W. M., and Martinez Lastra, J. L. (2022). Multimodal interface for human–robot collaboration. *Machines*, 10(10):957.

Rawal, N. and Stock-Homburg, R. M. (2022). Facial emotion expressions in human–robot interaction: a survey. *International Journal of Social Robotics*, 14(7):1583–1604.

Reddy, B. S. and Basir, O. A. (2010). Concept-based evidential reasoning for multimodal fusion in human–computer interaction. *Applied Soft Computing*, 10(2):567–577.

Rossi, S., Leone, E., Fiore, M., Finzi, A., and Cutugno, F. (2013). An extensible architecture for robust multimodal human-robot communication. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2208–2213.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Shani, R., Tal, S., Derakshan, N., Cohen, N., Enock, P. M., McNally, R. J., Mor, N., Daches, S., Williams, A. D., Yiend, J., Carlbring, P., Kuckertz, J. M., Yang, W., Reinecke, A., Beevers, C. G., Bunnell, B. E., Koster, E. H., Zilcha-Mano, S., and Okon-Singer, H. (2021). Personalized

cognitive training: Protocol for individual-level meta-analysis implementing machine learning methods. *Journal of Psychiatric Research*, 138:342–348.

Shu, B., Sziebig, G., and Pieters, R. (2019). Architecture for safe human-robot collaboration: Multi-modal communication in virtual reality for efficient task execution. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pages 2297–2302.

Shumanov, M. and Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior*, 117:106627.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Skantze, G., Hjalmarsson, A., and Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66.

Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7.

Thoker, F. M. and Gall, J. (2019). Cross-modal knowledge distillation for action recognition.

Tio, A. E. (2019). Face shape classification using inception v3.

Tulsiani, S. and Malik, J. (2014). Viewpoints and keypoints.

Wang, J., Tang, Z., Li, X., Yu, M., Fang, Q., and Liu, L. (2021). Cross-modal knowledge distillation method for automatic cued speech recognition.

Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X., Makris, S., and Chryssolouris, G. (2019). Symbiotic human-robot collaborative assembly. *CIRP Annals*, 68(2):701–726.

Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition.

Wilde, N., Kulić, D., and Smith, S. L. (2018). Learning user preferences in robot motion planning through interaction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 619–626.

Wongvibulsin, S., Frech, T. M., Chren, M.-M., and Tkaczyk, E. R. (2022). Expanding personalized, data-driven dermatology: Leveraging digital health technology and machine learning to improve patient outcomes. *JID Innovations*, page 100105.

Yi, D., Su, J., Liu, C., Quddus, M., and Chen, W.-H. (2019). A machine learning based personalized system for driving state recognition. *Transportation Research Part C: Emerging Technologies*, 105:241–261.

Zhao, M., Li, T., Alsheikh, M. A., Tian, Y., Zhao, H., Torralba, A., and Katabi, D. (2018). Through-wall human pose estimation using radio signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365.

Zhao, X. and Zhang, S. (2011). Facial expression recognition based on local binary patterns and kernel discriminant isomap. *Sensors (Basel, Switzerland)*, 11:9573–88.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FullvideoPFHRCom.mp4