Intelligent Supportive System for People with Profound Intellectual and Multiple Disabilities

Gašper Slapničar, Michał Kosiedowski, Arkadiusz Radziuk, Erik Dovgan, Torsten Hammann, Meike Engelhardt, Christin Kupitz, Carmen Campomanes-Alvarez, Dorota Janczewska, Peter Zentel, and Mitja Luštrek*

People with profound intellectual and multiple disabilities (PIMD) are a vulnerable and often marginalized group of people, commonly without the ability of symbolic communication. This makes expressing and recognizing their inner state and desires challenging and requires in-depth knowledge and understanding of each individual. Use of standard assistive technologies is thus infeasible due to lack of personalization. This challenge is tackled by the INSENSION system—a novel intelligent decision support system leveraging state-of-the-art noninvasive audio-visual sensor technologies together with machine learning algorithms and expert knowledge, to detect and interpret behaviors and communications (nonverbal signals—NVSs) of people with PIMD in challenging real-world scenarios. The detection of NVSs supports caregivers and allows the people with PIMD a form of communication through a communicator application. Furthermore, the system enables them to control and adjust their environment through a smart room and multimedia player. According to quantitative and qualitative evaluation, good recognition accuracy is achieved, and the system empowers people with PIMD and increases their quality of life.

1. Introduction

People with profound intellectual and multiple disabilities (PIMD) are characterized by below average intellectual functioning and adaptive behavior typically accompanied by sensory or physical impairments as well as complex health needs.[1,2] Their quality of life (QoL) is influenced by a complex interaction of personal and social environmental factors as described in the International Classification of Functioning, Disability and Health (ICF) by the World Health Organization (WHO). People with PIMD mostly communicate on a presymbolic level using unconventional behavior signals like specific vocalisations or body movements, which are meaningful to express their personal needs.[3] Therefore, these meaningful behaviors are highly individual, leading to a strong dependence on others in perceiving and adequately interpreting these communication attempts in all areas of life across the whole life span.[4] This high individuality complicates finding a common way of mutual understanding between people with PIMD and their environment.

Should the gap not be bridged and some level of understanding not attained, this group of people remains unfulfilled in terms of their wishes and feelings, which in turn negatively influences the satisfaction of their physiological and social needs, their cognitive, emotional, and communicative development, as well as general QoL.[5,6] At the same time, this puts a large burden on caregivers as well as the process is demanding, slow and sometimes frustrating.[7]

G. Slapničar, M. Luštrek
Department of Intelligent Systems
Jožef Stefan Institute Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: gasper.slapnicar@ijs.si

M. Kosiedowski, A. Radziuk
Poznan' Supercomputing and Networking Center
Jana Pawała II 10, 61-139 Poznań, Poland

E. Dovgan
BE-terna d.o.o.
Verovškova cesta 55a, 1000 Ljubljana, Slovenia

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202400925.

DOI: 10.1002/aisy.202400925

T. Hammann
Freiburg University of Education
Kunzenweg 21, 79117 Freiburg im Breisgau, Germany

M. Engelhardt, P. Zentel
Ludwig Maximilian University Munich
Geschwister-Scholl-Platz 1, 80539 München, Germany

C. Kupitz
University of Cologne
Herbert-Lewin-Straße 2, 50931 Köln, Germany

C. Campomanes-Alvarez
Data & Analytics
Seat Code Badajoz 97, 08018 Barcelona, Spain

D. Janczewska
Stowarzyszenie Na Tak
Chełmońskiego 2/1, 60-753 Poznań, Poland

In the age of omnipresent sensory and computational devices, the process of mutual understanding can be accelerated and enhanced with the use of technology;[8–10] however, people with PIMD cannot use existing interactive assistive technology (AT) due to their condition, as discussed in the following section.

1.1. Existing Assistive Technologies

Many ATs have been developed to enhance inclusion, accessibility to services, and the teaching-learning process for individuals with disabilities. These technologies aim to foster autonomy, independence, and the acquisition of social skills.[11] While complex digital ATs are utilized by individuals with visual, hearing, and physical impairments, those with intellectual or behavioral disorders, including autism spectrum disorder (ASD), are less likely to use such technologies.[11]

Common examples of digital ATs encompass Web 2.0 interactive applications, mobile learning platforms, adaptive digital boards, and custom wearable hardware. More specialized instances include robotics, which have been shown to enhance cognitive and social skills. Dor example, the NAO humanoid robot has been used to augment learning and communication skills in children with ASD, demonstrating promising outcomes for young individuals with severe intellectual disabilities.[12] Socially assistive robots (SARs) have been developed to perceive and respond to the emotional states of children with ASD, facilitating improved social interactions.[13] However, the implementation of such robotic technologies is often hindered by high costs and accessibility challenges.

In addition to robotics, VR applications compatible with commercial headsets have been developed to improve social interaction, communication barriers, and behavioral restrictions in individuals with disabilities.[14,15]

In the context of rare genetic disorders, such as KIF1A-associated neurological disorder (KAND), Angelman, Cornelia de Lange, Fragile X, and Rett syndromes, assistive technologies have been tailored to address specific needs and formidable challenges of such people. These include communication devices and personalized interventions that consider the unique cognitive and affective profiles of individuals with such conditions. Specifically, reinforcement learning is recently being utilized for continuous interaction with such people, adapting complexity, type of tasks, and gamification, to ensure user engagement and effectiveness towards rehabilitation goals.[16]

Many of these ATs, however, are not directly applicable for people with PIMD, as they require some level of symbolic understanding and independence. Instead, a more feasible alternative for such people are technologies relating to affective computing and mental state estimation[17] in order to determine what pleases or displeases them and to infer what they want. A large body of work again dealt with different levels of ASD—the most commonly investigated intellectual disability—where physiological signals (electroencephalogram—EEG, electrocardiogram—ECG, galvanic skin response—GSR, photoplethysmogram—PPG) were used to train deep learning (DL) systems to classify affective states of people with ASD.[18,19] Physiological data are most often obtained from wearable sensors and have been extensively used to recognize stress or emotional states such as happiness, sadness, anger, and fear, achieving significant accuracy.[20,21] More recent approaches propose contact-free alternatives for data capture to obtain rich physiological information in an unobtrusive manner via cameras and eye trackers.[22]

Another noninvasive alternative to physiological signals is audio, which can be captured unobtrusively. This is still challenging, as many people with disabilities are minimally or nonspeaking; however, nonverbal vocalizations are very valuable in terms of affective information. They were shown to allow for training of binary valence classifiers achieving average F1 scores above 0.7 in a leave-one-subject-out experiment, even with relatively few instances.[23]

Noninvasive ATs have overall been successfully presented to families of people with disabilities to enhance the understanding and treatment (especially in cases of ASD), offering more affordable and user-friendly technology that can complement conventional or robot-based therapy.[24]

Approaches relying exclusively on brain signals (EEG) are mostly encompassed under brain-computer interfaces and were also proposed for classification of emotions of people with disabilities.[25] Recent effort is also being put towards understanding the black-box DL algorithms for EEG-based emotion classification for people with ASD, by proposing novel interpretability methods such as RemOve-And-Retrain (ROAR).[26] Such methods support the recovery of highly relevant features from pretrained neural network, increasing trust and improving model understanding in this challenging domain.

Importantly, several datasets have been introduced and made available to support the development of systems for detection and analysis of affective states, tailored for people with disabilities. One example is CALMED (Children, Autism, Multimodal, Emotion, Detection), which provides multimodal data, including audio and video features, to aid in the creation of better affective computing applications and systems.[27] Another audio-focused example is ReCANVo (Real-World Communicative and Affective Nonverbal Vocalizations), which provides nonspeech vocalizations labeled by function from minimally speaking individuals, collected in real-world settings with inclusion of close family members. Such datasets are rare and highly valuable, facilitating development of novel ATs for people with disabilities.[28]

1.2. Unresolved Challenges and Our Contributions

As mentioned, many existing solutions heavily focus on ASD and often assume homogeneity of such people, limiting evaluation to existing datasets, often collected in controlled experiments. There is a distinct lack of work and more importantly practical real-world implementations dealing with heterogeneous PIMD people. Wu et al.[29] recently developed a contactless emergency assistance system designed for individuals with severe physical disabilities (tetraplegia without voice) in which they aimed to replace traditional emergency bells for emergency situations. This system confirms the trend of using contactless technologies for people with profound disabilities but is limited to a specific use case.

We can see that existing work in AT for people with intellectual disabilities only partially addresses the issues faced by people with PIMD, so we designed and implemented a holistic system, including sensoric setup, data processing and machine learning (ML) algorithms, and real-world validation. In this article, we give a complete overview of the proposed INSENSION system—a novel intelligent decision support system (IDSS) leveraging state-of-the-art noninvasive audio-visual sensor technologies together with ML algorithms to detect and interpret behaviors (or nonverbal signals—NVSs) and communication of people with PIMD in challenging real-world scenarios. INSENSION system was developed using a holistic interdisciplinary approach. Behavioral experts initially worked directly with caregivers to model the needs of people with PIMD (primary users) and their caregivers (secondary users). Computer vision (CV) and ML experts were then included in the loop to design a feasible unobtrusive sensing setup and develop ML models for detection of inner states and communication attempts. Contextual information was also considered during monitoring. Finally, the system executed applications manipulating the environment (e.g., stopped music, changed lighting, etc.) in an attempt to address the identified needs of primary users and increase their QoL.

The rest of this article is organized as follows. In Section 2, we give an overview of the INSENSION system design. We continue by giving detailed explanations of the developed intelligent methods in Section 3. The experiments and results are reported in Section 6. Finally, conclusions and discussion are given in Section 7.

2. System Design

We initially identified the base requirements of our system in order to achieve our goal of improving the QoL of people with PIMD. We started with the idea of a responsive environment in the context of people with disabilities and adapted existing validated pedagogical foundation relating user needs to required system functionalities.

2.1. User Needs

Our main goal was to provide a system that recognizes NVSs (comprising behavioral patterns), related them to the inner states and communication attempts of primary users, and intelligently responds. The categorization of possible inner states and communication attempts was proposed by experts in special needs education in collaboration with primary and secondary users. It resulted from prolonged observations of primary users in the early stages of the project and was based on theoretical foundations from literature.

Specifically, inner state was modeled on a Likert scale from 1 to 9, where 1–3 signify *displeasure*, 4-6 signify *neutral*, and 7-9 signify *pleasure*. This scale represents the whole range of emotional valence, which was recognizable in all people with PIMD involved in our research. It was inspired by existing validated scales specifically towards individuals with severe intellectual disabilities, such as the Disability Distress Assessment Tool (DisDat),[30] Mood, Interest, and Pleasure Questionnaire (MIPQ),[31] and methodologies outlined by Roemer et al.[32] The granularity of the scale was matched to the caregivers' ability to judge the primary users' inner state.

Three main communication attempts were defined—*demand*, *protest*, and *comment*. These were again inspired by existing validated scales, such as the Preverbal Communication Schedule (PVBCS).[33] They are recognizable in most people with PIMD and cover the majority of their communication—the desire for something (*demand*), its opposite (*protest*), and engaging in social interaction (*comment*).

Inner states and communication attempts of people with PIMD are more nuanced than described here, but the recognition of these nuances is challenging even for the caregivers, and they vary between people with PIMD.

Given the subtle differences between specific numeric labels and ambiguity of some behaviors, we hypothesized difficulty in the separation of such states and thus simplified the problem for ML by merging fine-grained numerical classes into broader categories.

These inner states and communication attempts are reflected mainly in audio-visual NVSs of the primary users, such as facial expressions, gestures, body movements, and sounds. Pedagogical experts and caregivers defined an extensive list of specific expressions and gestures, in part based on the literature,[30,32] which were known to be related to inner states or communication attempts of a specific individual (e.g., *widened eyes*, and *shaky body*). The caregivers then assessed the use of these NVSs by individual primary users involved in our research via questionnaires.[34] The caregivers were also invited to identify additional NVSs.

In addition to the paper-based assessment, the caregivers and experts in special needs education labelled inner states, communication attempts and NVSs in video recordings, specifically with the Elan software. The inner states and communication attempts were by necessity labelled subjectively, but both the caregivers and experts were involved to maximize the quality of the labels. Importantly, the assessment explicitly allowed for ambiguity—caregivers were asked to report the same behavioral signals in multiple affective states if applicable, to clarify ambiguous or context-dependent behaviors. This mixed-method approach (caregiver observation, structured questionnaires, video annotation) was used to ensure robust, reliable labeling of highly individualized communicative behaviors. The labelled NVSs were later refined to those that can feasibly be detected using state-of-the-art ML methods through RGB cameras and microphones.

In addition to the mentioned audio-visual NVSs, the physiological state of primary users was also expected to change alongside their inner state.[35] We thus monitored physiological parameters relating to heart rate (HR) and HR variability (HRV). These were computed via PPG, using both a wearable device as well as unobtrusively using only RGB camera recordings.

Furthermore, the interaction context is crucial to correctly interpret inner states and communication attempts, since the same NVS pattern can have a different meaning depending on the context. The latter was monitored using environmental sensors measuring brightness, loudness, and number of people. The identified contexts influencing meaningful behavior of primary users were grouped based on the effect they had on the primary user, which could be positive or negative. Example contexts eliciting positive responses included: swinging, listening to music, singing, massage device, watching movies, playing with toys, specific locations (e.g., outdoors), specific people (e.g., mother, father), specific sounds (e.g., birds), and changes in lighting. On the other hand, contexts eliciting negative responses included: changes in temperature (coldness), crowds, loneliness, sudden loud noises, unknown people, and changes in body position.

The specific assistive services to be provided were determined in close cooperation with secondary users, who best understand the primary users. The services were presented in the form of applications closely connected with the adjustable contexts described previously. The services included changes in noise/audio, lighting, suggested use of specific items/toys, and suggested changes in primary user body position. A specific service was invoked based on the contextualized inner state or communication attempt detection, and suggestions were given to secondary users alongside the detections. This allowed secondary users, who were caregiver experts most familiar with each primary user, to also evaluate the system in action and provide feedback, allowing the system to improve via active learning.

2.2. System Architecture

To address the requirements described in the previous section and move from conceptual foundation towards practical implementation, we proposed the system architecture shown in **Figure 1**.

At a high level, the proposed system comprises the following components: data inputs for the IDSS at the bottom (blue), the IDSS back-end system (orange), and the front-end applications with which users interact (red).

At the first level of the system, the facial expression, identity, and gesture recognizer worked with two RGB camera video inputs collected from different angles and aimed to recognize facial expressions and gestures of the primary users, alongside identification of primary and secondary users. Physiological signals were the inputs from the Empatica E4 wristband, specifically the PPG. Early attempts included remote estimation of physiological signals using RGB video to minimize obtrusiveness of the sensor setup,[36] but due to challenging real-world conditions, it was later decided that a wearable was more reliable. Vocalization recognizer aimed to recognize specific relevant voices and sounds captured with microphones. Object recognizer again used RGB camera video to detect relevant objects predetermined to influence the primary users (e.g., favorite toys). Finally, context sensors provided information about the environment, such as lighting conditions, temperature, and humidity.

Once these data were prepared and processed, they were fed to the machine learning IDSS and enhanced with detected contextual information to produce a contextualized decision, which was in turn executed via applications. The latter then influenced the environment directly to respond to the NVSs of primary users (smart room, media player). Caregivers could also evaluate the suggestions and help improve the model via active learning.

We describe each component in more detail in the following sections.

3. Intelligent Monitoring Methods

We first describe the NVS recognition models, which are the individual data inputs for the IDSS and comprise several state-of-the-art ML models working with sensor inputs, mainly video and audio. The NVSs encompass behavioral patterns and communicative signals, such as vocalizations, facial expressions, and gestures. We will then discuss the IDSS backend, comprising the context recognition models and ML models for predicting inner state and communication attempt based

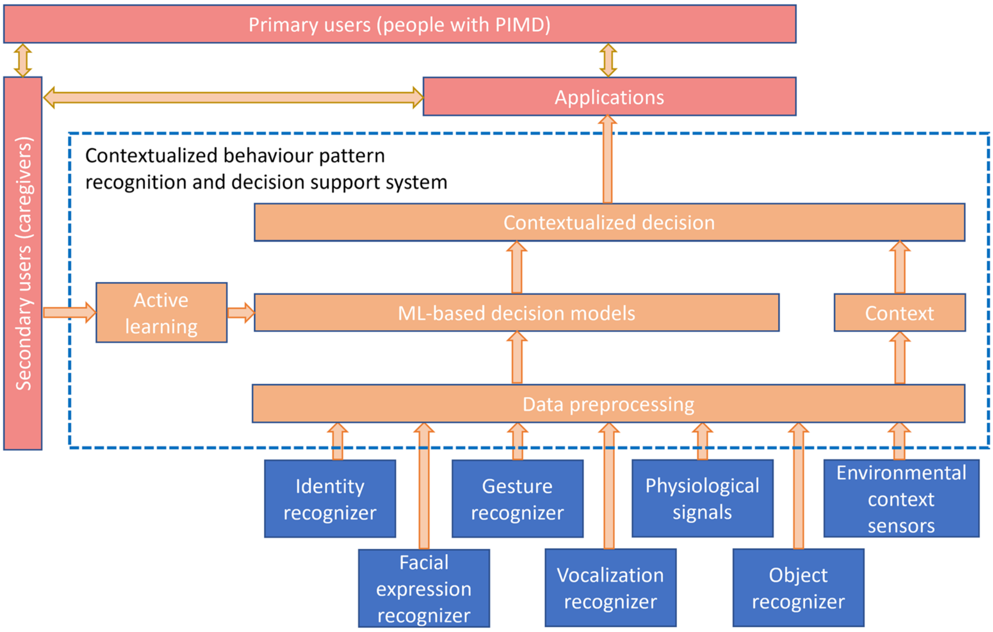

Figure 1. The architecture of the proposed system. Blue are the data inputs for the decision support system, orange is the decision support back-end system, and red are the front-end systems alongside users.

on the previously observed NVSs. Our pipeline also includes active learning for model improvement.

3.1. Nonverbal Signal Recognition

Models for classification of each class described in this section were trained on the data obtained in the first phase of the project data collection, lasting several months and including 6 primary users. Details are reported in **Table 1**.

In cases of data shortage for some specific gestures, additional data was collected from other people who imitated specific gestures. While the number of primary users is small, it should be noted that it is exceptionally difficult to obtain data from people with PIMD, as confirmed by very limited literature on such people in the context of intelligent assistive technologies. People with PIMD are rare and very sensitive to any changes in their daily environment and routine, so any obtained data is very valuable.

In terms of specific sensors, our data collection setup included two Logitech C920 Pro HD cameras, recording at 1080p, two Rode VideoMicro microphones with 100 Hz–20 kHz frequency range, a Raspberry PI4-based custom sensing component with an integrated Bosch Sensortec BME680 sensor for air quality (including temperature, humidity and air pressure), and a SparkFun TSL2561 luminosity sensor with a range of 0.1–40 000 Lux. Subjects were also equipped with an Empatica E4 wristband, capturing a variety of physiological signals. The devices were mounted in elevated positions, perpendicular to one another, placed on maneuverable extensions allowing for precise position selection. The aim of this elevated placement was to capture as much audio-visual information as possible, having good subject exposure and less occlusion. An example setup with all the sensors at one of the premises is shown in **Figure 2**.

3.1.1. *Identity Recognizer*

This component was based on video from RGB cameras, and it was added due to the need to identify and exclusively monitor the movements, gestures, vocalizations and physiological parameters of the primary user, as well as identify other important people (e.g., secondary users, and parents).

To do this, we performed face detection, face encoding/embedding, and person identification. Firstly, the histogram of oriented gradients (HOG) descriptor[37] was used to detect all the faces that appear in a video frame. These were then cropped and passed to the FaceNet model,[38] which produced facial embeddings. Based on the distances between these embeddings, a threshold method and traditional ML classifiers were used to recognize the primary user and other people for which we had reference images, as well as discard other people in the scene. We investigated eight well-established algorithms for classification of specific individuals with PIMD: naive Bayes, k-nearest neighbors (kNN), support vector machines (SVMs), logistic regression, stochastic gradient descent (SGD), simple fully connected neural networks (NNs), random forest (RF), and eXtreme Gradient Boosting (XGB).

Table 1. Data collected in the first phase of the project to train recognizers in preparation of the pilot study.

	Participant X1	Participant X2	Participant X3	Participant X4	Participant X8	Participant X9
Sex	M	M	F	M	M	M
Age [y]	9	7	18	9	3	16
Recorded sessions	7	14	13	34	9	11
Total time [h]	03:04:37	06:27:20	04:00:30	10:04:57	04:05:09	04:38:30
Neutral/none inner state instances	511	717	1407	3540	1081	779
Pleasure inner state instances	342	930	827	530	1279	2058
Displeasure inner state instances	134	28	156	410	98	78

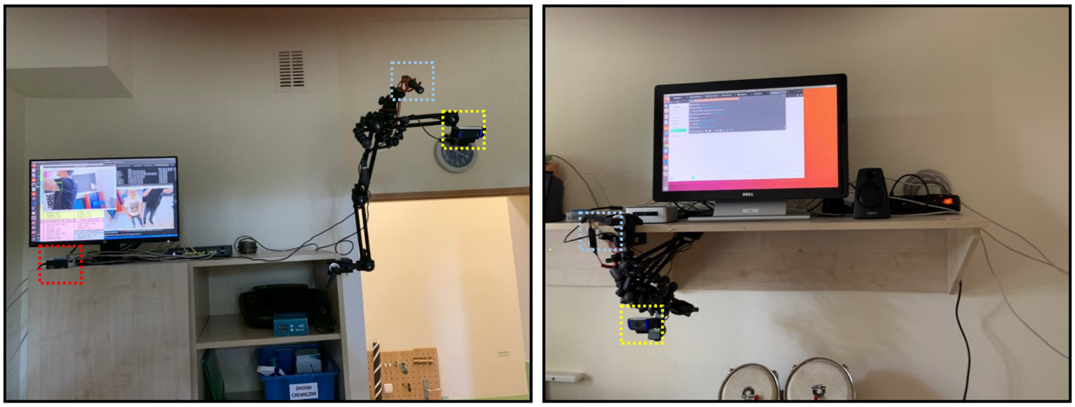

Figure 2. An example sensor setup at one of the premises. Yellow rectangles are cameras, blue are microphones, and red is the Raspberry PI4-basd sensing component.

Additionally, a cropped image of the face obtained in this stage was also needed for facial expression recognition discussed in the next subsection.

3.1.2. Facial Expression Recognizer

The recognizer was designed to recognize the facial expressions indicative of inner state or communication attempts, which were identified as described in Section 2.1. The general approach was based on the use of the facial landmark positions of the identified primary user in each frame of the video. The extraction of these landmarks was made using OpenPose,[39] which is able to provide facial landmarks and body keypoints in real time, as shown on the example in **Figure 3**. A standardization strategy of centering and scaling was used in order to assure the robustness to zooming, camera proximity and subject movements, as well as to facilitate the performance in people with different body sizes.

An LSTM-based classification model was developed for the jaw appearance recognition. In particular, the system was able to detect the jaw-related classes of *neutral*, *biting*, *grinding*, and *drooping*. To train this model, a combination of data from the INSENSION project and a subject providing this set of expressions was labelled and then used to train an LSTM network. The inputs for jaw appearance were composed of the distances between the aligned landmarks of the chin and the mouth, for a sequence of 40 consecutive frames. First, the distance between the top-mid point of the mouth (obtained with OpenPose) and the bottom-mid point was calculated. Then, the difference between the × coordinates and the y coordinates of selected points of the mouth and the chin, and the bottom-mid point were obtained. In total, the input vector was formed by concatenating seven computed values for a sequence of 40 frames.

The facial expression recognizer consisted of further classifiers with the same underlying idea as described for the jaw appearance. Concerning the mouth, four different classes were of interest: *corners up* (smiling), *corners down* (frowning), *wide open*, and *lip movements*. For this purpose, another LSTM network was trained, achieving good performance.

The system recognized four different classes of eye appearance, i.e., *closed*, *semi-closed*, *widened*, and *winking*, by calculating the distances between specific eye-related landmarks and using hand-crafted rules. Furthermore, an LSTM network has been trained to detect the movements of eyebrows, resulting in detection of *frown* and *raised* expressions.

The nose region was the most challenging due to the inconsistent landmark detection by OpenPose in this part of the face. Some movement of the nose keypoints implies changes in the lateral wings of the nose, which was not consistently captured by the system. Nevertheless, an LSTM-based classifier was proposed to deal with the nose movement recognition.

Different people composed the train and test datasets in order to assure that the experiments with facial expressions were person-independent. More specifically, the dataset has been split, for each facial expression, into a training dataset and a testing dataset, in such a way that the samples for four randomly selected people were used to train and the samples for the remaining one (a different person) were used to test. We experimented with network configurations comprising 2–6 LSTM layers and a softmax output layer. The number of neurons per layer depended on the size of the input. The input size ranged from 50 (5 values in 10 consecutive frames for the movements of eyebrows and mouth) to 280 (7 values in 40 consecutive frames for the slower-changing jaw appearance). The widest internal layers had three times the number of input neurons (150–840) and the narrowest a quarter of the input (12–70). The best-performing configuration was always chosen to be used.

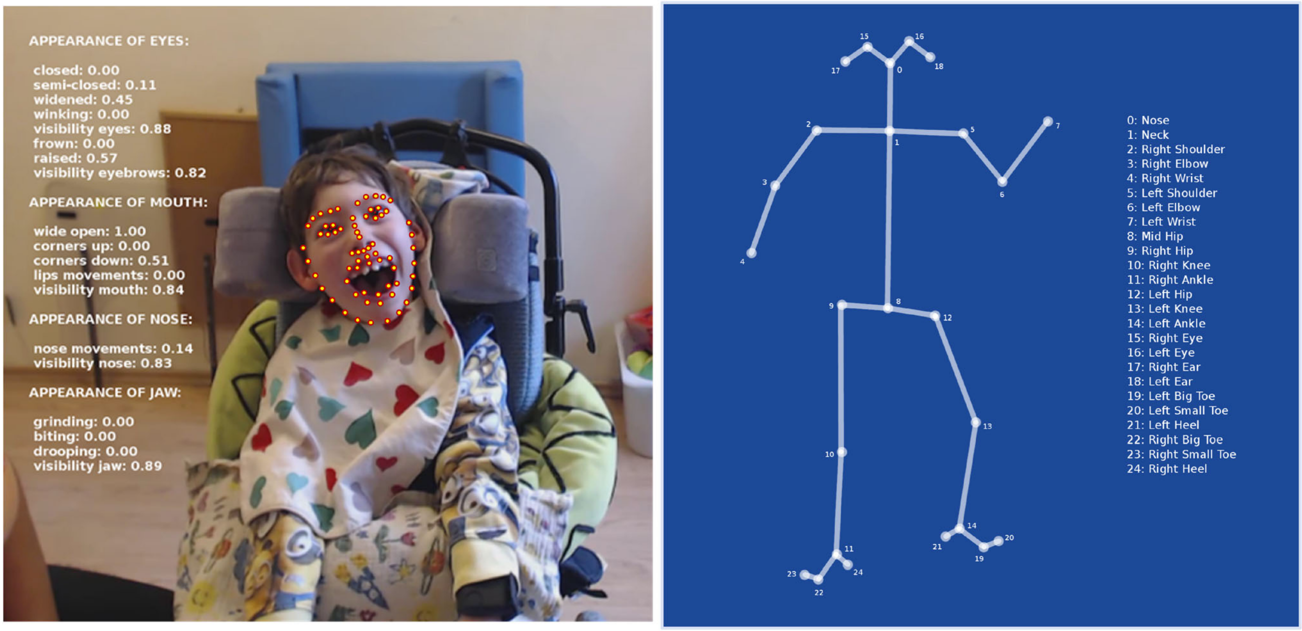

Figure 3. Facial and body keypoints detected by OpenPose[39] and used for facial expression and body gesture recognition.

3.1.3. Gesture Recognizer

This component aimed to detect the body poses and motions for each primary user. The following groups of poses and gestures were investigated: body posture (*jerky* and *leaning*), appearance of the head (*floppy, shaking, nodding, raised, turned,* and *leaning*), appearance of each arm (*rigid, floppy, jerky, stretched, flexed, raised,* and *close to body*), appearance of the hands (*hand on hand* and *hand on head*), appearance of each leg (*stretched, flexed, raised,* and *rubbing*), and appearance of the feet (*foot on foot*). Similarly to other design choices, this chosen set was based on a combination of existing literature, pedagogical experts and caregivers who knew these children the best. A set of poses and gestures that was determined to be informative and clear by the focus groups consisting of caregivers and pedagogical experts was used. Detections were again based on keypoints detected in each frame of the videos using the DL OpenPose framework, as shown in Figure 3.

In practice it often happens that keypoint detections have low probability or are simply not detected due to occlusions. In such cases, the whole frame was discarded for subsequent pose and gesture estimation. While this limits responsiveness and generalization of the system in real-world scenarios, we focused on clear, unambiguous, and useful detections for the recognizers and IDSS. This decision leads to cleaner data coming to the IDSS and increases trust in the system by the secondary users (caregivers), which is important. As before, the keypoints were also standardized, including scaling and centering.

Some gestures can only be detected through several frames in which locations of the keypoints change, so the gestures were split into static poses and dynamic actions. The former can be inferred from a single frame and the latter requires a sequence of frames without missing data. Traditional ML classification models (random forest, SVMs, logistic regression, and k-nearest neighbors) based on keypoints were developed for gestures that do not have clear anatomical rules (jerky movement of the body, head movements, jerky movement of the arms, and legs rubbing). On the other hand, for the remaining cases, rules were designed by following the reasoning about the anatomical working of the human body joints.

For example, in the case of the pose related to the flexibility of the arm, the approach consists of calculating the angle between the shoulder, the elbow, and the wrist. Thus, functions were defined to establish a value between 0 and 1 for each pose (*flexed* or *outstretched* in this example) depending on the angle. A threshold was then experimentally determined to separate such classes and finalize the rule. For gestures that change in time, the angle change was temporally tracked and a pattern was determined that defined a specific movement (e.g., *raising arm*).

Moreover, for poses that change in 3D space, especially those relating to the face, the camera keypoint coordinates had to be mapped between 3D world coordinates using existing coordinates and camera properties (focal length, optical center, and radial distortion). The procedure we used is called direct linear transform (DLT) and is illustrated in **Figure 4** alongside a real-world example of 3D head movements. For details on DLT, we refer the reader to [40].

3.1.4. Vocalization Recognizer

Vocalizations of primary users cannot be mapped to “words” of any language and might not have any distinct sequences of phonemes as is typical for people without PIMD. This means that traditional automatic speech recognition (ASR) methods are not feasible. Instead, two main methods of vocalization recognition were considered: hidden Markov models (HMM) supported with Gaussian mixture models (GMM) and neural network-based (NN) classifiers. Both methods took well-established Mel-frequency cepstral coefficients (MFFC) representation of recorded audio as input vectors. These were computed using a sliding window where different lengths between 50 and 1000 ms were investigated. These lengths were determined based on shortest detected vocalization length. The feature vector was then extended by the 1st and 2nd order derivatives of the MFFC vector and by its maximum autocorrelation coefficient. The labelled vocalizations for initial models took 6 possible values in total, across two subjects (*laughs, snores, wails, squeals, hums,* and *vocalizes*). The labelling was done by caregivers as described in Section 2.1. Only vocalizations that occurred frequently enough were considered.

In the HMM approach, every unique vocalization type was stored as a list of distinct states of the vocalization event and the state transition matrix. Each state was assumed to correspond to one stationary segment of audio observations (a segment that was expected to appear within a modeled event); the stationary signal in a given state was thus represented by its GMM that describes distribution of features mentioned before. The training procedure included two phases. The first unsupervised audio frame clustering was conducted using the GMM method, followed by HMM parameter estimation until convergence using expectation-maximization (EM) method. The NN-based approach investigated recurrent neural networks (RNN), specifically LSTMs with different architectures (number of layers and neurons). In the end an LSTM with 5 layers, each containing 100 neurons was chosen as the best-performing model. This was developed later as an alternative to the HMM + GMM approach and was evaluated on an updated set of audio recordings, again belonging to two subjects. The subjects were new because the two from the earlier data collection were not available. The subjects' availability was an issue throughout our research because people with PIMD experience frequent health problems. Moreover, the data was collected during the COVID-19 pandemic, resulting in longer periods of unavailability due to parents taking measures to lower the risk of their children getting in contact with the virus. The new recordings were labelled with 9 possible values in total (*laughs, coughs, grunts, moans, aaa, eee, aeaeae, eeh,* and *nge*). The labels differed from earlier ones in part because vocalizations are highly individual and the two subjects in this data collection voiced different ones and in part because they were labelled more specifically in an attempt to capture more information. The instances were still short windows between 50 and 1000 ms, and the audio in each window was represented using MFFCs.

Moreover, the vocalization recognizer[41] was extended with two additional components. First was called ambient sound recognizer and it used the same LSTM architecture, which was

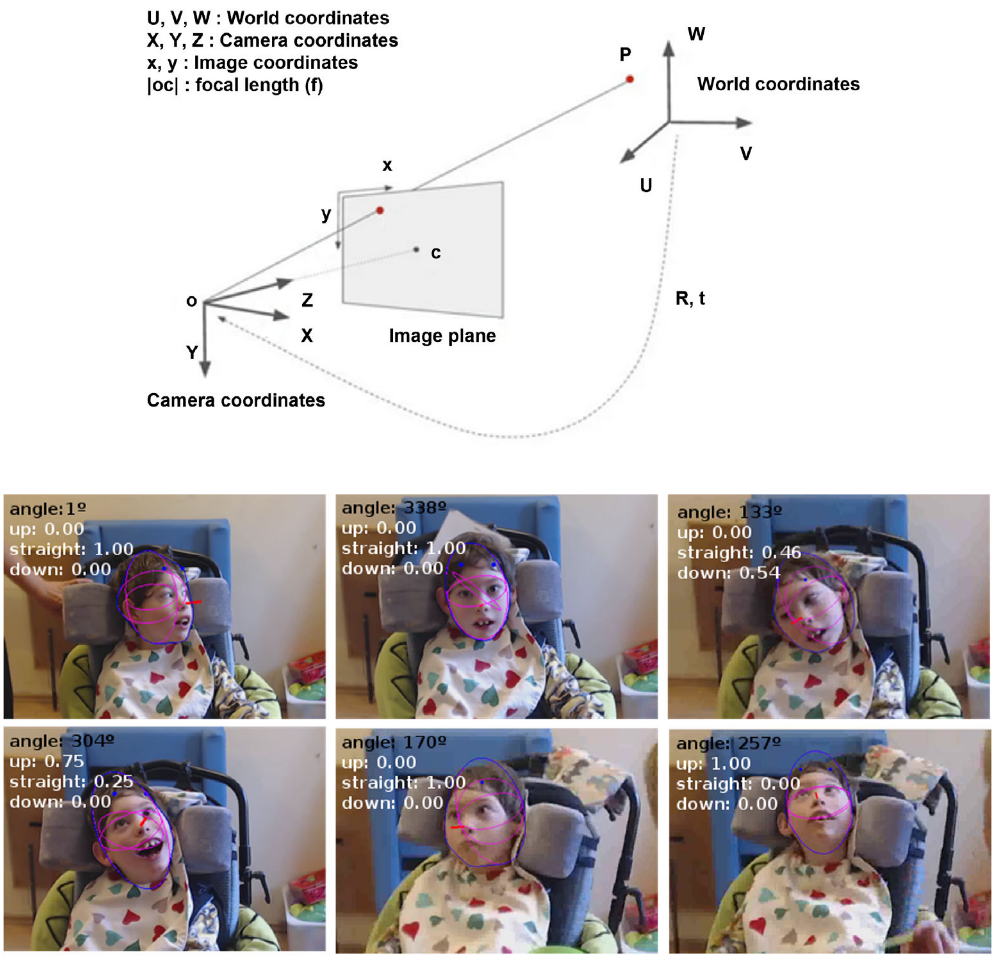

Figure 4. 3D coordinate mapping and real-world example. Ellipses relate to pitch, roll, and yaw of the head.

trained to detect 9 ambient sound classes: *ambulance fire brigade*, *animals*, *massager*, *music instrument sounds*, *singing*, *toy noises*, *vehicles*, *violin*, and *background*. Data was augmented by multiplying original instances and imposing different noise on the original recordings and the model was trained and evaluated in a 5-fold cross validation experiment using different window lengths between 50 and 3000 ms. The second component was used for detecting sudden loud noises in the environment. We used a simple root mean square of the audio signal, which further amplified extreme amplitude variations corresponding to sudden loud noises. Detector was based on average loudness at a given time, compared to preceding period, and was used to signal occurrence of loud sounds.

Vocalization classification has seen large progress since the project, especially relating to disabilities. Updates to the proposed methodologies via state-of-the-art DL methods should be considered. A recently proposed multilevel fusion of wav2vec2, mel-spectrograms, and other descriptors showed impressive performance across 7 vocalization classes from the open-access ReCANVo dataset,[42] outperforming traditional approaches used in INSENSION. We believe building on these ideas and incorporating such datasets into foundation models for audio-based affective state classification is a promising direction.

The individuality of the vocalizations requires adaptation of the recognizer to each primary user, making vocalization recognition more challenging than recognition of facial expressions or gestures. The ambient sound classes were also defined based on the environment and toys the specific individuals used and/or encountered often (e.g., during play time or physical therapy). This information originated from the caregivers with experience and knowledge of specific individuals. So, again, should a new user or rather a new environment be included, or existing environment changed, this would require updates of corresponding recognizers.

3.1.5. Physiological Signals

As briefly mentioned previously, we initially attempted contact-free estimation of physiological signals using remote PPG (rPPG); however, such methods are still difficult to implement in challenging real-world conditions with inconsistent lighting and exposure, which was apparent in our early results.[36] We thus decided to use the PPG obtained with an Empatica E4 wristband, as it was more stable and reliable. Empatica was worn on the wrist in accordance with device guidelines. Some participants initially had trouble tolerating a wearable tightly equipped to their wrist; however, acceptance was relatively high after the introduction period. Importantly, there were some physical challenges in regards to equipping the wristband, as many users were exceptionally thin and fragile given their condition. It was sometimes challenging to ensure consistently good connection between the sensors and the skin, inevitably leading to some low-quality data, which we tried to salvage through preprocessing.

The obtained PPG was first band-pass filtered to remove movement noise. It was then segmented into standard 30-s windows, which are long enough to capture the periodic cardiac activity and any physiological changes caused by a change in affective state. The latter do not usually change very often and abruptly. Accordingly, the majority label for each window was considered as ground truth. In the next step, the systolic peaks in the PPG signal were determined using a state-of-the-art derivative-based algorithm.[43] Differences between subsequent peak locations were then used to obtain HRV, which was in turn used to compute features widely used in literature, such as standard deviation of peak-to-peak intervals (SDNN), root mean square of successive peak differences (RMSSD), and standard deviation of successive peak differences (SDSD).[35]

The computed features were then used to classify the inner states and communications attempts defined in Section 2.1, using standard ML classifiers (k-nearest neighbors, random forest, eXtreme Gradient Boosting, SVMs, and AdaBoost) each with a set of optimized hyperparameters obtained through 5-fold CV grid search. We then performed 5-fold CV experiments predicting inner state or communication attempt (in separate experiments), monitoring standard classification metrics. Additional details on features, models, and evaluation are reported in.[35] These physiology-based classifications were in turn fed to the IDSS as additional information to produce a decision.

3.2. Context Recognition

Contextual information was obtained from video (interaction with specific objects or people) and ambient sensors (noise levels, temperature, and lighting conditions). Relevant contexts for each primary user were provided via questionnaires that were filled by the secondary users and pedagogical experts. The identified (interaction) contexts and known NVSs were modelled and stored using an ontology, which contained a codification of our limited world. We chose the Protégé framework[44]—an open-source ontology editor and knowledge management framework developed by Stanford University, widely used for building and visualizing semantic ontologies through formal representations to present this ontology. In our assistive monitoring context, we had several classes representing the main entities present in the system, including *PIMD user*, *Sensors*, and *Applications*, alongside corresponding interactions between them. A graph of interactions in the ontology is shown in **Figure 5**; however, precise details are omitted for brevity. We mention it as an important stepping stone leading to identification of relevant contexts described in the following sections.

To facilitate recognition of relevant contexts listed in Section 2.1, we relied on previously described components, which allow for identification of persons and objects of interest and their interaction within the environment. The proposed ontology also served as guidance for formalization of expert knowledge described in the next section.

3.2.1. Object Recognizer

A set of objects that were known to elicit behavioral changes in primary users was provided by secondary users. Examples of such objects included *guitar*, *vibrating toy*, and *mug*. A number of ground-truth images of these objects were made from several perspectives in order to finetune a neural network for object recognition.

To recognize these objects in each frame of the video streams, we used the pretrained CNN-based You Only Look Once (YOLO) architecture,[45] which we finetuned with the provided images of objects of interest. We also performed standard data augmentation (rotations, scaling, cropping) to increase the number of learning examples. The network is relatively light-weight and performs well in real time, making it suitable for a complex system with high computational requirements.

Given recent progress in CV, especially with the emergence of foundation models (e.g., SAM2) and Vision Transformers, we believe such fine-tuning on specific objects might no longer be necessary and an out-of-the-box solution will be usable for most objects and environments.

3.2.2. Environmental Context Sensors

While some contextual information could be obtained from previously described components, the NVSs of primary users can also be importantly influenced by environmental conditions. For instance, the primary user can close their eyes either due to the room being too bright or due to them being frightened, which is an important distinction when modelling their inner state. The monitored set of such environmental contexts included luminosity level, air temperature, and air humidity. These were monitored with a dedicated RaspberryPi microcomputer equipped with suitable sensors for each of the parameters.

4. Contextualized NVS Detection and Decision Support System

The core of this component of the INSENSION system was the developed IDSS, whose purpose was to use all the other behavioral and context data to recognize the inner state and communication attempt of the primary user. We initially investigated traditional ML classification methods. In later stages we developed a new approach that instead used rule sets based on ML and formalized expert knowledge.

4.1. Traditional Machine Learning

The first approach implemented several traditional ML algorithms (linear discriminant analysis—LDA, *k*-nearest neighbors—kNN,

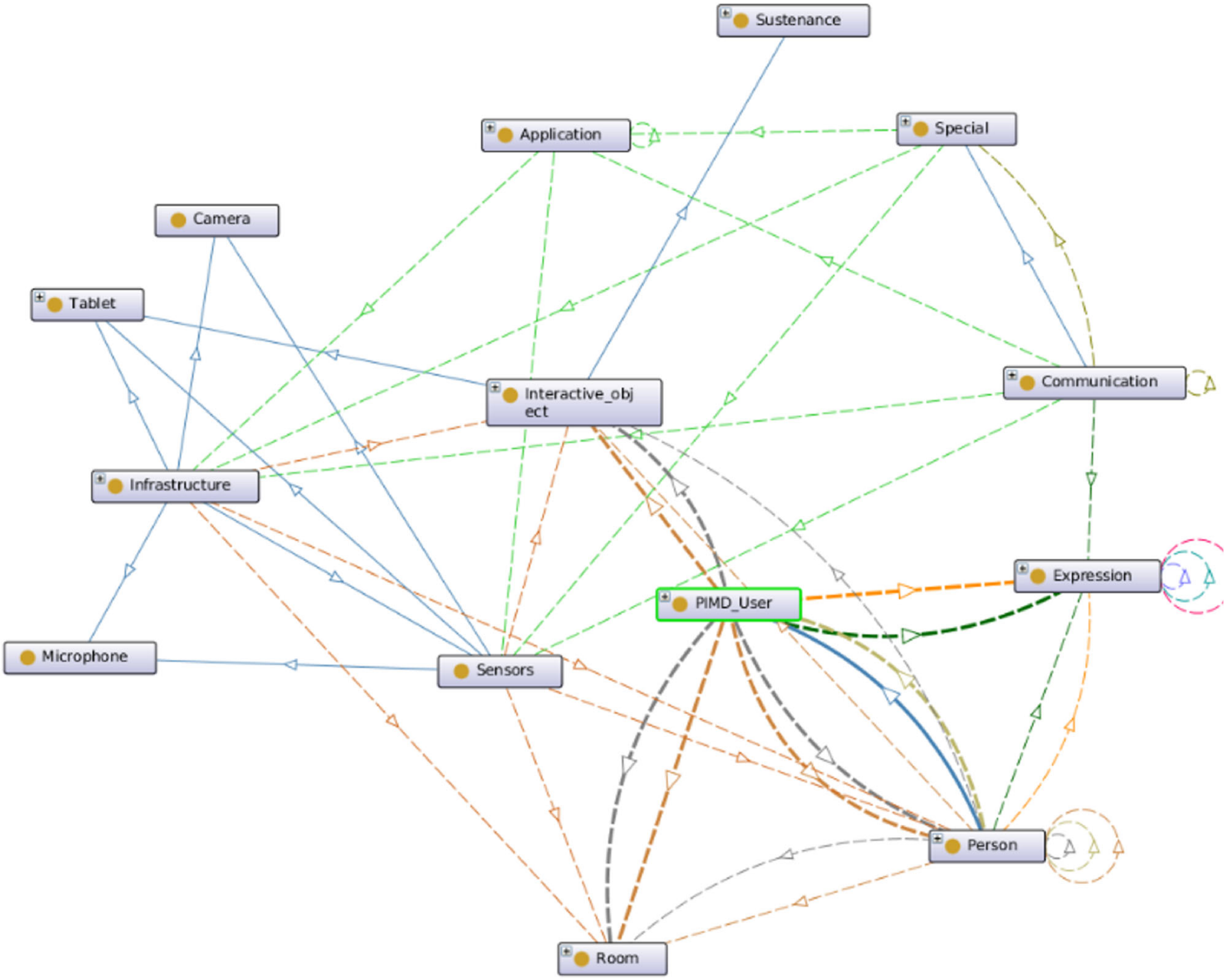

Figure 5. A graph of interactions between classes in our ontology.

support vector machines—SVMs, Bayes, decision tree, and random forest). All models were trained using the previously described facial, gesture, physiological, and vocalization features, which were obtained as averages in a 10-s sliding window of each video. Specifically, these features included the three following main groups: a) facial expressions, gestures, and postures, directly based on the outputs of the recognizers described in Section 3.1; b) vocalizations corresponding to individual vocalizations relevant for each user; and c) physiological signals represented by a range of HRV features, together with inner states and communications attempts recognized from physiological signals only. The feature set could be extended with the detected contextual information, but we limited ourselves to primary user information as the context information (such as which caregivers or objects are present) is likely to be specific to the situations in our recordings and thus not suitable for models that should be useful in a broader range of contexts (e.g., primary user moved to new location or new secondary user present).

Furthermore, we added some derived features, such as average of all values above 0 and histograms of behavioral values in a window—for each basic feature, we obtained 10 histogram and histogram density features. If the inputs were not present (no detection in a window), we imputed the data using decision tree regressor models. In total, this brought the size of our feature set to over 100 (depending on the number of recognized ambient sounds, which varied based on location), so we decided to reduce the feature set using correlation analysis, removing highly correlated features with Pearson's correlation coefficient over 0.7. Additionally, we also investigated hyperparameter tuning for the best-performing model.

We removed instances without detections for facial expressions, gestures and postures—which were deemed the most important group based on experts and caregivers—and obtained the distributions reported in **Table 2** and **3**.

We can see that the data was quite imbalanced in terms of inner state and communication attempt labels. This is mostly because the objective of caregivers is to prevent displeasure, protest and to some degree demand, and it would not be ethical to elicit these inner states and communication attempts on purpose. We resolved this using two methods: a) random oversampling,

Table 2. Distribution of instances in regards to inner states.

Inner state	All instances		Instances with physiological data	
	User A	User B	User A	User B
Neutral	561	307	291	307
Pleasure	102	154	18	154
Displeasure	82	/	27	/

Table 3. Distribution of instances in regards to communication attempts.

Comm. attempt	All instances		Instances with physiological data	
	User A	User B	User A	User B
None (unrecognized)	648	433	319	433
Comment	79	9	13	9
Demand	18	14	4	14
Protest	/	5	/	5

which randomly selects instances of the smaller classes and generates copies of them until all classes are balanced, and b) synthetic minority oversampling technique (SMOTE),[46] which is an established better-performing method for oversampling. Balancing was always done on training data only, so it could not happen that one copy of an instance would be in the training data and another in the test data.

4.2. Expert Knowledge and Rules

To obtain expert knowledge, we initially encoded data from questionnaires filled by secondary users, which contained the connections between NVSs and inner states or communication attempts. Rules were derived from a simple rule template defined by the experts. According to the template, when a set of NVSs are observed for a particular user, their inner state or communication attempt belongs to a particular class. A snippet showing an example of this initial expert knowledge is given on left side of **Figure 6**.

Upon subsequent inspection and evaluation of the initially proposed rules, it was determined that the initial set is too rigid to describe the reality well, as some rules did not always hold true or required ad-hoc modifications to accurately reflect the behaviors of primary users. We thus redesigned the rule template to be more flexible. According to the template, each rule can have several conditions, and the class into which it classifies. Each condition is defined with the attribute, i.e., the NVS that is observed, the threshold that has to be met to trigger the condition, and the condition weight. The thresholds are applied to probabilistic outputs of the recognizers, which correspond to how strongly a NVS is expressed and how confidently it is recognized. The weights determine how important a condition is for a particular rule.

The expert model then classified an instance as follows: for each condition that was met, i.e., the attribute exceeded the threshold, the condition's weight was returned; otherwise, 0 was returned. The returned values were summed up for each rule into a rule weight. The rule weights of all rules in the model were summed up for each class value independently. In addition, the maximum rule weights, i.e., the weights of rules in the case when all conditions were met, were also summed up for each class value. The maximum rule weights were then used to normalize the actual rule weights (per class) as follows.

$$w_{c,\text{norm}} = \frac{\tan^{-1} w_c}{\tan^{-1} w_{c,\max}} \tag{1}$$

where w_c is the sum of rule weights per class and $w_{c,\max}$ is the sum of maximum rule weights per class. At this point, we had normalized rule weight per class that corresponded to the evidence we had for each class. Next, the normalized value was weighted with class weight and compared to the class threshold. If the weighted normalized value did not exceed the threshold, it was set to 0. This was done for each class independently, to allow for the possibility that some classes require more evidence to be recognized than others. Finally, the class with the highest weighted normalized value was assigned to the instance.

The experts could provide correct relations between NVSs and inner states and communication attempts (i.e., the attributes and classes), but they found it difficult to think in terms of exact probabilities that a certain NVS will occur and be recognized within a certain window (i.e., they could not provide the weights and thresholds). These weights and thresholds were therefore optimized computationally using the differential evolution algorithm. This is a stochastic optimization method inspired by biological evolution. It starts with a population of randomly generated solutions (vectors of weights and thresholds in our case). Each solution mutates with a certain probability, which means that the difference between two randomly selected solutions is added to it (this is where the name *differential* evolution comes from; the rationale is that the magnitude of the mutation is the same as the magnitude of differences between solutions).

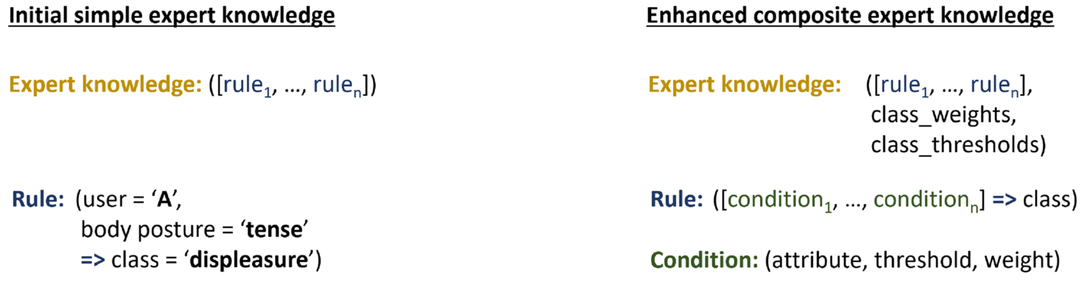
Initial simple expert knowledge
Expert knowledge: ([rule1, …, rulen])
Rule: (user = 'A',
body posture = 'tense'
=> class = 'displeasure')
Enhanced composite expert knowledge
Expert knowledge: ([rule1, …, rulen],
class_weights,
class_thresholds)
Rule: ([condition1, …, conditionn] => class)
Condition: (attribute, threshold, weight)

Figure 6. A partial example of extracted expert knowledge. Initial attempt on the left side, enhanced data-driven rule templates on the right side.

The mutation is followed by crossover, meaning that pairs of solutions are combined into new ones—their offspring. Mutation and crossover yield a number of new solutions, which are evaluated and compared with the previous ones, and the best are selected for the next generation. The whole procedure is repeated for a predefined number of generations, after which the optimization is concluded and the best solution returned.

The evaluation of a solution in the context of the differential evolution was conducted as follows. The weights and thresholds comprising the solution vector were used to complete the expert ruleset. This ruleset was then applied to annotated recordings, and the balanced accuracy was used as the evaluation metric for the solution. Like traditional machine learning, expert rules were evaluated using crossvalidation (CV; see Section 6.3). The optimization was treated as training, so the evaluation that was part of the optimization was always performed on the training portion of the data only.

4.3. Decision Fusion and Ensembles

While we initially planned to fuse expert and machine-learned rule-based models, we eventually opted for decision fusion instead of model fusion. This means that we fused decisions of individual models using an ensemble approach. There were two reasons for this: on one hand, it proved difficult to develop a well-performing method for rule fusion, while on the other hand, ensembles are an established, flexible and typically well-performing approach.

Ensembles were built by combining the best machine learning and expert models mentioned previously. More precisely, machine learning and expert models were tested with various test settings in respect to *Primary User*, *Class label*, and *Set of attributes* (three possible sets: a) facial expressions, gestures and postures, b) vocalizations, and c) physiological signals).

For each of the tested settings, the previously mentioned six ML models (LDA, naive Bayes, SVM, kNN, DT, and default RF) in addition to expert system and a hyperparameter-optimized RF were built (eight models in total). We used fine-grained information for inner state and communication attempt classes, where each label could take integer value on a scale from 0 to 10, with the middle value indicating perfect neutrality. Afterwards, the models were sorted according to their balanced accuracies. Based on this, the ensembles were created by combining different models based on performance, confidence and type (ML or expert knowledge).

4.4. Active Learning

The idea of active learning is that if a NVS is erroneously recognized, caregivers may provide a correction, which is used to label the respective instance, this instance is then added to the training data, and models are retrained to improve the accuracy in the future. We devised six active-learning strategies differing in the instances that are added to the training data (all or only some of the incorrectly classified ones, or also some correctly classified ones) and the weight they are assigned (more weight to the newly added instances or not). Specifically, the strategies differed as follows. 1) **Strategy 1:** All incorrectly classified instances are labelled and added to the training data. 2) **Strategy 2:** All incorrectly classified instances are labelled and added to the training data in five copies. The idea is that corrections are emphasized in the hope that the models are improved faster. 3) **Strategy 3:** Randomly selected 50% of the incorrectly classified instances are labelled and added to the training data. This simulates the caregiver not having the time or inclination to correct all mistakes. 4) **Strategy 4:** Randomly selected 50% of the incorrectly classified instances are labelled and added to the training data in five copies. 5) **Strategy 5:** Randomly selected 50% of all instances (correctly or incorrectly classified) are added to the training data. The idea is that even new instances that can already be recognized correctly can improve the decision models. 6) **Strategy 6:** Randomly selected 50% of all instances are added to the training data in five copies.

We tested the active learning by not using all labelled data for training initial models, and using the rest to simulate caregivers' corrections.

Active learning improved the accuracy of NVS recognition compared to the initial models, however the final models were not quite as accurate as models built from all the data, because none of the active-learning strategies assumed the caregivers would label all newly added instances.

5. Assistive Applications for Users

The previous section described detection of NVSs from the features characterizing the user, i.e., facial expressions, gestures, and vocalizations. However, the decisions to be made to improve the users' QoL typically involve the context of the users. Therefore, in order to make meaningful decisions, we firstly need to determine the relation between detected NVS and the context, which we did using a statistical approach that asserts the similarity between the context data and the target class. First, for each context feature, a contingency table (i.e., a table of frequency distribution of the variables) between the feature and NVS was computed. Next, we assessed the amount of relationship between the feature and NVS using the chi-square test for independence. We selected the maximal amount between the one obtained with inner state and the one obtained with communication attempt. Finally, the context features were sorted based on amount of relationship in descending order, which determined their rank.

This data-based approach is combined with expert-knowledge-based approach in which experts define the relevance of the context. They defined the order of the context features for each user, similar to the output of the statistical approach. As the result, two ordered list of features were stored in the knowledge base for each user. They were then combined by assigning to each context feature the highest rank among the two lists.

IDSS was designed to interpret the NVSs of people with PIMD as their inner state or communication attempt and estimate the probability of context features as causes of these behaviors. While this is already valuable on its own, our system also provided a mechanism that allowed it to undertake actions based on the identified understanding of the behaviors—assistive applications. These were designed in such a way that they allowed the caregivers to define rules that influence decisions of the assistive applications concerning what action to undertake in what kind of situation. Such functionality was achieved by introducing rule-based decision making and prioritization lists of identified causes within individual applications, which could be modified by the secondary users. This is summarized in **Figure 7**.

The available assistive applications in the system included the communication application or communicator, multimedia player, and smart room application.

5.1. Communicator

The purpose of this application was to translate what a person with PIMD is experiencing (inner state) or attempting to communicate (communication attempt) into messages understood by others. This especially concerns those that do not know the particular person with disability well (e.g., new caregivers). The communicator messages also relate to the context of the monitored primary user, which may include for example the presence of other people in the room or actions performed by other applications (e.g., a song played by the media player). The messages to secondary or tertiary users are provided in natural language via a graphical user interface (GUI).

The communicator additionally contained configurations about each individual user, allowing for high personalization. Each configuration consisted of a set of rules, which are the basis for the decision of what message to show, given the detected inner state or communication attempt and context. The secondary users can edit these configurations and rules. They can also provide feedback on the detection and suggested action, in accordance with active learning paradigm.

5.2. Multimedia Player

The main rationale for the multimedia player was to facilitate independence of the primary users, keeping them happy without direct involvement of a second party. One of the envisioned ways was by focusing their attention on music and other sounds. Such an autonomous service would empower the primary users, which was hypothesized to cause short-term (satisfaction) and long-term (sense of influence on the neighborhood) positive effects. Furthermore, this would relieve some burden from the secondary users (not having to spend all their time next to the primary user), which is seen as beneficial to the relationship between the two.

The situations in which the media content was played again depended on the configuration of the application entered by the secondary user. This person was considered the expert in understanding what multimedia could be played (i.e., are liked or potentially liked by the given primary user), in response to which NVSs and within what context. They prepared playlists and rules based on time of day and the recognized inner state or communication attempt.

5.3. Smart Room Application

The goal of this application was to enable direct interaction between the electrical devices in the room and the person with PIMD. To achieve this, this application provides means for controlling electrical devices, through transforming detected NVSs into specific actions, again according to a user-specific configuration in the application. The control of devices included switching electrical or virtual devices on or off. Electrical devices included a fan, a lamp, or a vibrating mattress (often used to stimulate people with PIMD). Example of interaction via lighting is shown in **Figure 8**.

6. Experimental Evaluation

It is important to note that while we conducted a pilot study in the final stage of the project, individual component evaluation was done at different earlier phases to validate their performance. Following these individual results, we also provide system performance overview on the pilot data. Importantly, despite low number of involved subjects, such rich multimodal data of people with PIMD is immensely valuable but difficult to collect. The results are highly subject-specific, but still offer important insights into the feasibility of such a complex system.

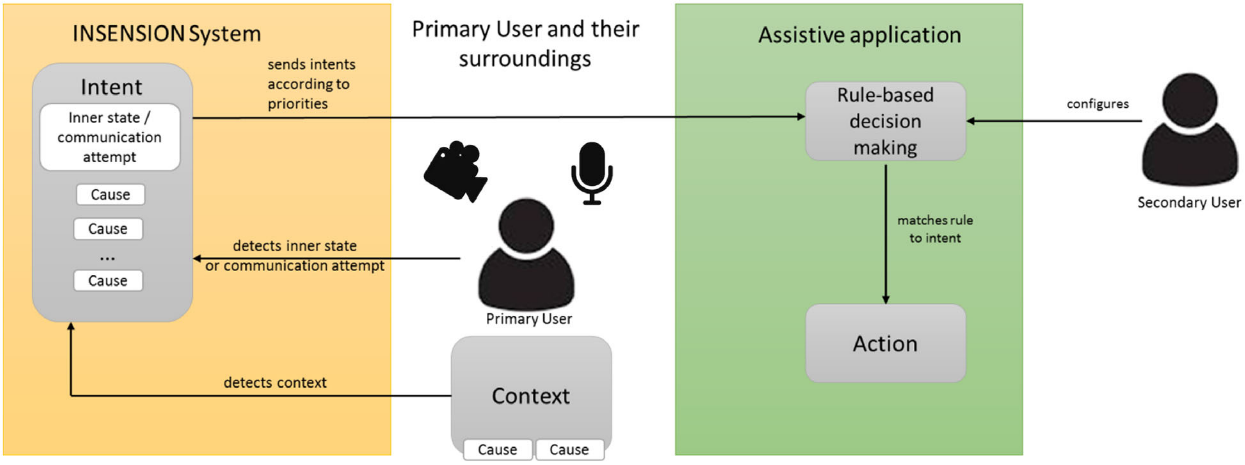
INSENSION System
Intent
Inner state / communication attempt
Cause
Cause
...
Cause
sends intents according to priorities
detects inner state or communication attempt
detects context
Primary User and their surroundings
Primary User
Context
Cause
Cause
Assistive application
Rule-based decision making
matches rule to intent
Action
configures
Secondary User

Figure 7. The role of assistive applications in the proposed system.

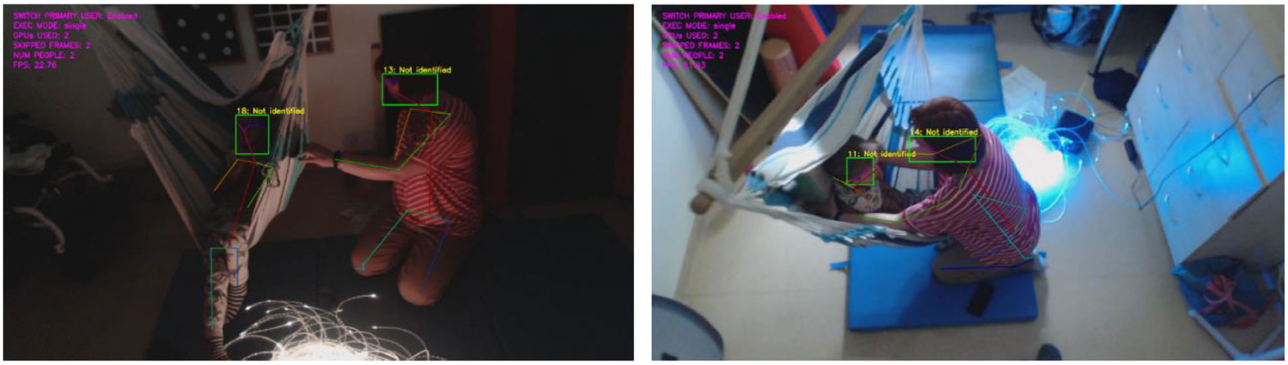

Figure 8. Interaction between the primary user and the smart room application via lighting.

The methodology of this study was oriented toward a single-case design (SCD) approach,[47] which is the most common way to deal with the high heterogeneity of the target group and the resulting small sample size.[48] Current research on people with PIMD uses SCD as methodology following the trend towards the increased integration of technology. Inter alia, several studies focused on the investigation of the emotional expression by analyzing physiological data with machine learning algorithms[35] or visual inspection,[49] the support of communication by classifying the personal movements and environmental data with machine learning algorithms (i.e., different data regarding locations, times or weather),[50,51] the enhancement of attention, body movements and affective behavior using an interactive technical ball,[52] or the promotion of physical activity through technology aids.[53,54]

6.1. Primary Users and Ethics

The six individuals with PIMD participating in the project had different specific conditions, including but not limited to Sturge-Weber syndrome, cerebral palsy, epilepsy, hypotonic form, quadriplegia, cytomegalovirus disease, and post-inflammatory hydrocephalus. Four primary users had visual impairment and two had hearing impairment. Five subjects were male and one was female, with ages ranging from 3 to 18 years (mean = 10.9 y). During the duration of the project, participation intensity of each primary user varied, depending in their condition and agreement of their legal representatives. In some early evaluations, we only used more plentiful data of two individuals, which was continuously collected through several recording sessions. However, in the final pilot, 6 primary users participated. This is the reason for differences in the data used for evaluation of different components at different stages of the project.

Given exceptional sensitivity of the primary users and their data collected in the INSENSION system, ethical clearance was obtained before the project commenced. Importantly, informed consents were also obtained from legal representatives of each participant (most commonly parents) about their participation in the project and subsequent use of data for model training and project dissemination. In adherence with the consents, the collected data was safeguarded and used only for model training and internal validation of the system. It was not distributed in any way outside of the project, except specific images that were allowed to be used for scientific dissemination (e.g., Figure 3,4).

6.2. Evaluation of Individual Components

Individual components were evaluated offline with post analysis. This meant that data was recorded and then annotated after the recording took place, which was followed by individual evaluations described in this section.

6.2.1. Identity Recognizer

The ML models were evaluated on a dataset containing frames obtained with two RGB cameras at 1920 × 1080 resolution. In total, it consisted of 3000 facial images of 6 individuals, split evenly among participants. The dataset was split into five subdatasets of 100 images per person for a 5-fold CV approach. The images of each subdataset have been extracted from different videos to assure the variability and to avoid overfitting. Thus, each fold was composed of 400 images per person (2400 in total) for training, and 100 images per person (600 in total) for testing.

Furthermore, two more datasets were constructed for people who are not primary users. The first included Related-Unknown People (RUP), like relatives and caregivers, which included images of 15 people. The second included General-Unknown People (GUP) completely out of the scope of the project, which included images of 395 people.

We monitored standard classification performance metrics, including accuracy and F1 score. For classification of the primary users (identification), all models achieved similar performance within 3%, with SVM achieving the highest accuracy and F1 score of 99.2%. It similarly showed best performance in separating known and unknown people, while having one of the lowest execution times due to using a linear kernel.

6.2.2. Facial Expression Recognizer

As highlighted in Section 3, most expressions were recognized using ML techniques, while eye appearance was classified based on anatomical rules. 1) *Jaw appearance (4 class values: grinding, biting, drooling, and no movement)*. For this case, we used the data of 5 subjects who had these class values and performed a LOSO experiment again. The best-performing LSTM achieved average accuracy and F1 of 80%. 2) *Mouth appearance (5 class values: lip movement, corners of mouth up, corners of mouth down, mouth wide open, and neutral)*. Similar to the previous case, we used the collected data of people with PIMD, augmented with the CK+ dataset. We again split it into 70% for training and 30% for testing, where the best LSTM model achieved 88% accuracy and 87% F1 score. 3) *Eye appearance (5 class values: closed, semi-closed, widened, winking, and neutral)*. We used an anatomical model based on distances between eye keypoints. For training this models, we used a dataset containing annotated recordings of people with PIMD from the project (minority), as well as additional videos where 5 people performed expressions of interest from three points of view (frontal, semi-lateral, and lateral). The dataset was split into a training and a testing subset in a leave-one-subject-out (LOSO) manner. This allowed for a subject-independent evaluation, which achieved average accuracy and F1 score of 89%. 4) *Eyebrow movement (3 class values: frown, raised, and no movement)*. We again conducted a similar LOSO experiment for the eyebrow movement, where the best-performing LSTM model trained on sequences of 10 frames achieved average accuracy and F1 score of 80% and 78%, respectively. 4) *Nose movement (2 class values: movement and no movement)*. Concerning the nose region, we augmented the limited data collected from people with PIMD with the Extended Cohn-Kanade Dataset (CK+).[55] Due to subtle movements of the nose, we opted for a binary classification. This dataset was split into training and testing using the 70–30% split, not following the LOSO schema, since each instance in the CK+ dataset corresponds to a different individual. The best-performing LSTM model here achieved 73% accuracy and F1 score.

For the evaluation of facial expression and gesture recognizers, both PIMD and non-PIMD data was used for training. The latter filled the gaps for specific expressions or gestures that were under-represented in the former. This was further alleviated using SMOTE oversampling.[46] In the final evaluation, only PIMD data was always used in the test set.

6.2.3. Gesture Recognizer

Similar to facial expressions, we also had two types of models for gestures—those based on ML techniques and those based on anatomical rules. 1) *Body posture (2 class values: jerky and leaning to the side)*. For the case of leaning, we used an anatomical rule by computing the angle between the hips-axis and the shoulder-axis. We again used a dataset comprised of collected data from people with PIMD and people purposefully doing these postures. Using the mentioned rule, we achieved accuracy and F1 score of 84% and 81%, when doing binary classification between leaning and no leaning. For the jerkiness, we instead trained ML models, which were fed sequences of keypoint characteristics as instances. A simple fully connected neural network performed best with accuracy and F1 score of 81%. 2) *Appearance of the head (7 class values: floppy, shaking, nodding, raising, turns to side, leans to side, and neutral)*. The first task included movements (shaking, raising, turning, and nodding). As this was not consistently present in collected data, we only used the data purposefully created by people without PIMD. A LOSO experiment was conducted, in which RF achieved the best average results of 89% accuracy and F1 score. The second task included nodding, which is characterized by alternative up and down movement of facial keypoints. Using a simple anatomical rule, we achieved accuracy of 98% and F1 score of 87% in the same LOSO experiment. Next, for the leaning of the head, we computed the angle between the eyes-axis and the shoulder-axis, obtaining accuracy of 72% and F1 score of 73%. Finally, the floppy head was also formed as a binary problem, using a more complex set of anatomical rules related to 3D positions as described in Section 3. We trained the model on people without PIMD performing the action and validated it on data of people with PIMD, achieving accuracy and F1 score of 86%. 3) *Appearance of joints (arms, hands, legs, and feet)*. For brevity, we report the results of all joint appearances combined. These were mostly (with the exception of jerkiness and hand rubbing) detected using anatomical rules, with accuracies and F1 scores between 79 and 98%. For jerkiness, a fully connected neural network performed best with accuracy of 72% and F1 score of 71%.

6.2.4. Vocalization Recognizer

We used a dataset of two primary users A and B, for whom we had PCM 16bit 16 kHz mono audio. As briefly mentioned in Section 3, the best-performing model was the 5-layer LSTM. It was evaluated in a robust temporal 5-fold CV experiment, where each recording was split into 5 subsequent parts with the aim of minimizing overfitting. Instances of four parts were used for training and the instances of the fifth part for testing. The shortest window length of 50 ms (for each instance) achieved the best results. Using this window length, the LSTM model achieved accuracy and F1 score of 72% for subject A and 87% for subject B, when classifying 9 possible vocalizations (*aaa*, *eee*, *aeaeae*, *eeh*, *nge*, *grunt*, *moan*, *laugh*, and *cough*).

It should be noted that data augmentation was also investigated, by multiplying instances of minority class and adding different noise. However, augmentation efforts yielded a worsening of results by 2–3% on average, so it was ultimately not used.

Additional component of the vocalization recognizer—the ambient sound recognizer—was also evaluated using the same 5-fold CV experiment. The LSTM-based ambient sound recognizer achieved highest accuracy and F1 score of 0.91 and 0.90, using a window length of 1000 ms, when classifying 9 groups of sounds: *ambulance fire brigade*, *animals*, *massager*, *music instrument sounds*, *singing*, *toy noises*, *vehicles*, *violin*, and *background*. The loud noise detector was calibrated on each environment-specific audio data and a threshold was determined each time to signal a loud sound.

6.3. Inner State and Communication Attempt Recognition

Due to some evidence in literature on connection between cardiovascular physiological parameters and inner states,[35,56] we initially attempted to classify inner states and communications attempts of people with PIMD using physiological signals only. We used the data of two subjects, which totaled 15 recording sessions. We split the PPG signal obtained from the Empatica wristband into 30 s windows and computed HR and HRV related features based on systolic peak locations. The data instances were stacked temporally (to avoid overfitting) and a 5-fold CV experiment was done. Six ML algorithms were compared (kNN, DT, RF, SVM, AdaBoost, and XGB), and the best results were obtained using the XGB classifier. It produced accuracy of 62% and F1 score of 59% for the inner state, and 48% and 45% for communication attempt, surpassing the baseline majority classifier by 12% on average.

This simple early approach was later contrasted by the sophisticated IDSS detailed in Section 3, which used the outputs of all other components to predict inner state and communication attempt. We used 5-fold CV for evaluation per-subject and per-class. We used all the available data of 2 primary users for evaluation. Initial observations showed that using SMOTE oversampling consistently outperformed random or no oversampling in all experiments. It was additionally found that audio-visual sets of features (from cameras and microphones) consistently outperformed physiological features, as the latter were unreliable due to problems in signal quality (the subjects are often jerky and have very thin wrists, making the PPG from wearable corrupted). However, physiological signals can in some ensemble cases help improve the results, depending on subject and ensemble, as seen in **Table 4**. The results of individual models and ensembles are shown in Table 4. Note that while we used balancing methods (SMOTE) on training data, we did not balance the test data. Considering that our test set classes remain imbalanced, we used balanced accuracy as the evaluation metric. Balanced accuracy is the average recall across all the classes, and recall is the fraction of instances belonging to a class that are in fact recognized as such.

Table 4. Summary of average balanced accuracies for different users and predicted classes. We always used SMOTE on training set and used all audio-visual inputs. *bestN* indicates the *N* best performing individual ML models used in the ensemble. Best3 and best2 cases included the 3 and 2 individually best-performing ML algorithms from those compared—these can be seen in the upper part of this table.

	Subject A inner state	Subject A comm. attempt	Subject B inner state	Subject B comm. attempt
Expert system (ES)	48.4%	33.3%	60.8%	35.0%
LDA	56.3%	73.3%	63.5%	**62.3%**
Naive Bayes	53.5%	58.0%	64.8%	44.8%
SVM	60.2%	57.0%	67.1%	48.8%
kNN	57.9%	51.1%	59.1%	44.8%
DT	54.0%	56.9%	59.5%	59.0%
RF	49.3%	52.4%	62.5%	35.0%
RF optimized (RFO)	59.8%	70.5%	65.5%	60.8%
best3	62.1%	72.5%	**69.3%**	53.6%
best3 + physio.	58.2%	75.3%	67.2%	56.0%
ES + best2	**62.4%**	72.2%	**69.3%**	50.6%
ES + best2 + physio.	59.2%	**76.4%**	66.9%	0.0%
RFO + best2	62.1%	72.5%	67.3%	53.6%
RFO + best2 + physio.	58.2%	75.3%	66.2%	56.0%
RFO + ES + best1	60.6%	72.2%	67.0%	50.6%
RFO + ES + best1 + physio.	59.0%	**76.4%**	66.9%	56.4%

We can observe that different models or ensembles perform the best for different subject and class. Despite these variations, the trend shows that most best results appear in the lower part of Table 4, indicating that ensembles tend to outperform individual models. This is however not the case for subject B when classifying the communication attempt.

6.4. Pilot Evaluation of Assistive Applications

After the system was developed and offline evaluation was conducted on per-component level, we also tested it live in a pilot involving six primary users (people with PIMD) and eight secondary users (caregivers) to evaluate the correctness of the decisions and their usefulness. We wanted to test each assistive application twice per primary user; however, we only managed to obtain partial results due to health-related absences. Specific scenarios were designed to elicit certain states and validate the behavior and influence of the applications in those specific situations, which were assessed by the secondary users.

The caregivers were asked for feedback through a survey where they were asked "How do you rate the following app features?" for the first four features in **Figure 9** and "Please answer to what extent you agree with the following statements" for the last three features.

The users were generally satisfied with the design of the user interface, as evidenced by high scores in language comprehension, transparency and visual readability, navigation, aesthetics, and general ease of use. While the user interface was not as refined as mature commercial software, the users felt that the more mundane aspects were executed well, and most challenges were due to the fact that our system was attempting to perform a novel and difficult task. There was substantial variability in their views on whether the application mirrored communication methods of other known messengers ($n = 4$ with positive reviews). While being able to converse with people with PIMD through a messenger-style application would be desirable, the purpose of these applications is bidirectional communication, which is not what our system does. So, perhaps the survey question was not entirely appropriate. Most critically, the idea of integrating the system into daily routines received divided opinions ($n = 3$ with positive reviews). Some concerns were expressed over the time required for setup. Considering the hardware used and the need for personalization, complex setup is inevitable, although some streamlining is possible. We believe, however, that most secondary users would accept it the payoff was sufficient, i.e., if the system was helpful enough. The most serious concerns were over system reliability. Given the task of the system, it can never work perfectly, although improvements are possible and certainly something to strive for in the future. However, the fact that half of the (admittedly small group of) caregivers could imagine using our system on a daily basis, even though it was far from a mature product, can be considered a success.

A summary of the key results from the perspective of interaction between primary users and the INSENSION system is given in **Table 5**. The Summary column compares the primary user's inner states as annotated by the secondary users and the

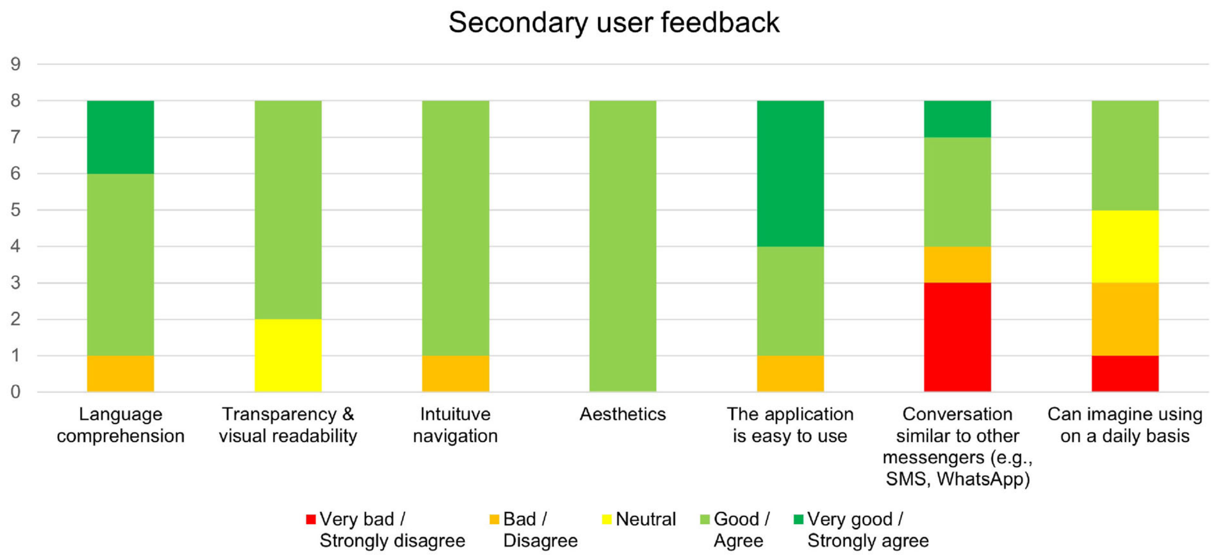

Figure 9. Secondary user feedback on the INSENSION application.

Table 5. Key findings in terms of interaction between primary users and the INSENSION system.

User	Scenario/app	Summary	Evaluation
X1	Smart room	Annotated neutral state. IDSS detected displeasure, triggered the application and then detected pleasure.	OK (incorrect but harmless detection, improved inner state if detection of pleasure was correct)
X2	Media player	Annotated displeasure. IDSS did not detect it, so application was not triggered.	Bad
X2	Smart room	Annotated displeasure (two cases). IDSS detected one displeasure and triggered the application.	Good
X4	Smart room	Annotated displeasure. IDSS detected the displeasure, but also other states, and triggered the application several times. It is possible this contributed to subsequent pleasure.	OK (some unnecessary triggering of the application)
X8	Media player	Annotated displeasure. IDSS did not detect it, so the application was not triggered.	Bad
X8	Smart room	Annotated displeasure (several cases). IDSS detected most cases and triggered the application.	Good
X8	Media player	Annotated neutral and pleasure. IDSS detected several cases of displeasure and triggered the application. It is possible this contributed to subsequent pleasure.	OK (incorrect but harmless detection, improved inner state if detection of pleasure was correct)
X8	Smart room	Annotated displeasure (two cases). IDSS detected the displeasure, but also other states, and triggered the application several times. After some triggers it detected pleasure.	Good (triggering until pleasure is achieved)
X9	Media player	Annotated neutral. IDSS detected neutral and pleasure, did not trigger the application	Good
X9	Smart room	Annotated neutral. IDSS detected neutral and pleasure, did not trigger the application	Good

detections of the INSENSION system. It also states the actions taken by the system based on its detections. The Evaluation column provides our evaluation of the system's performance. The evaluation is *good* when the system detected the inner state with a reasonable accuracy and acted appropriately. It is *bad* when the system made a wrong detection and did not act appropriately. It is *OK* when the system made an incorrect detection or did otherwise not act as intended, but the result for the primary user still appeared to be good.

From a pedagogical perspective, Table 5 shows that the IDSS has potential to support people with PIMD by adapting to their unique needs and improving their emotional well-being, as seen in cases like X2 (Smart room) and X8 (Smart room). However, inconsistent detection, such as in X2 and X8 (Media player), can lead to frustration and hinder development, highlighting the need for reliable responses. Accurate feedback, like in X9, promotes emotional regulation and autonomy, while unnecessary triggers, as in X4, can create confusion. It should be noted that sometimes even incorrect detections by the IDSS can lead to good results, such as in the case of X1 and X8 (Media Player). While these successes cannot be attributed to IDSS, they can be attributed to the system as a whole, since most of the actions it was designed to do are pleasurable for the primary users in most circumstances. This fortunately gives the IDSS some leeway in its challenging tasks. Overall, the system requires refinement and should be implemented for a longer period of time to be able to make reliable statements about the effects of the applications on the behavior of the participants as well as their development and QoL.

7. Conclusion

We proposed, developed, and validated a system for holistic support of people with PIMD. Using an interdisciplinary approach, we incorporated knowledge of pedagogical experts and caregivers to model inner states of primary users and relate them to their subject-specific NVSs. The latter were detected using several modules based on CV, audio processing, physiological signal analysis, and environmental context. These multimodal signals were used to train an IDSS, which detected inner states and communication attempts of primary users, and responded via applications influencing their smart room environment.

The results obtained in the INSENSION project, being an important novel step in proposing a technological solution supporting interaction of people with PIMD with their surroundings, yielded a deep insight into solving this challenge. The pilot trial confirmed that the developed platform achieves the desired outcomes to an extent. Using the system in some of the test scenarios increased the primary users' chance of receiving the appropriate support. This was especially noticeable for those scenarios in which the person with PIMD stayed alone and was supported only by the system. Individual components also worked in line with results of individual evaluations reported earlier. Importantly, this study's design was based on the assumption that the person with PIMD understands to some degree that the external circumstances (i.e., interaction with the caregiver, music played, and function of the device started) are or can be changed by their behavior. However, especially with this target group, it can be assumed that it would take more time for the primary users to fully grasp their ability to interact with the environment via the system.

An important challenge that remains is fluidity of behavior depending on context and environment. For instance, should the environment or caregivers change, the NVSs of PIMD individuals might change as well—either completely or more subtly, only in the meaning behind the same NVS.[57] Capturing such changes would require recalibration and retraining of the system based on new contexts.

INSENSION system is a pioneering approach in holistic unobtrusive AT for people with PIMD, who are incapable of symbolic communication and thus cannot use existing solutions. It represents an important stepping stone towards enabling such people to interact with their environment using responsive technology, which can ultimately increase their QoL.

Acknowledgements

The research presented in this article has been conducted within the INSENSION project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 780819. We would additionally like to acknowledge the support of the Jožef Stefan Institute (JSI) and the Slovenian Research and Innovation Agency (ARIS). Finally, we would like to acknowledge the work of Waldemar Jęśko, who importantly contributed to the audio analysis.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data collected in the project are not available due to the sensitive nature of both the data and the subjects involved. Permissions were granted for use of data in the project and related publications, but not for full data dissemination.

Keywords

computer vision, intelligent decision support, machine learning, physiological monitoring, profound intellectual and multiple disabilities, sensors

Received: October 29, 2024
Revised: June 24, 2025
Published online:

[1] H. Nakken, C. Vlaskamp, *J. Policy Pract. Intellect. Disabil.* **2007**, *4*, 83.
[2] K. Petry, B. Maes, C. Vlaskamp, *J.Appl. Res. Intellect. Disabil.* **2005**, *18*, 35.
[3] V. Munde, P. Zentel, In *Oxford Research Encyclopedia Of Education*. Oxford University Press, **2020**.
[4] A. K. Axelsson, J. Wilder, *Int. J. Dev. Disabil.* **2014**, *60*, 13.
[5] A. K. Axelsson, C. Imms, J. Wilder, *Disabil. Rehabil.* **2014**, *36*, 2169.
[6] B. Maes, G. Lambrechts, I. Hostyn, K. Petry, *J. Intellect. Dev. Disabil.* **2007**, *32*, 163.
[7] M. Engelhardt, *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete* **2021**, *2024*.
[8] P. Washington, H. Kalantarian, J. Kent, A. Husic, A. Kline, E. Leblanc, C. Hou, O. C. Mutlu, K. Dunlap, Y. Penev, M. Varma, *JMIR pediatr. Parent.* **2022**, *5*, e26760.
[9] M. Engelhardt, M. Kosiedowski, I. Duszyńska, *J. Enabling Technol.* **2020**, *14*, 87.
[10] M. Kosiedowski, A. Radziuk, P. Szymaniak, W. Kapsa, T. Rajtar, M. Stroinski, C. Campomanes-Alvarez, B. R. Campomanes-Alvarez, M. Lustrek, M. Cigale, E. Dovgan, G. Slapničar, *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conf. (IntelliSys)*, London, UK, Springer **2020**, Vol. *2*, pp. 895–914.
[11] J. M. Fernández-Batanero, M. Montenegro-Rueda, J. Fernández-Cerero, I. García-Martínez, *Educ. Technol. Res. Dev.* **2022**, *70*, 1911.
[12] M. A. Saleh, F. A. Hanapiah, H. Hashim, *Disabil. Rehabil.: Assist. Technol.* **2021**, *16*, 580.
[13] S. Cano, C. S. González, R. M. Gil-Iranzo, S. Albiol-Pérez, *Sensors* **2021**, *21*, 5166.
[14] V. Bravou, D. Oikonomidou, A. S. Drigas, *Retos: nuevas tendencias en educación fsica, deporte y recreación* **2022**, *45*, 779.
[15] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, M. J. Matarić, *ACM Transactions on Human-Robot Interaction (THRI)* **2022**, *11*, 1–28.
[16] F. Stasolla, K. Akbar, A. Passaro, M. Dragone, M. Di Gioia, A. Zullo, *Front. Psychol.* **2024**, *15*, 1372769.
[17] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, N. B. Hussin, *Comput. Educ.* **2019**, *142*, 103649.
[18] L. Muñoz-Saavedra, F. Luna-Perejón, J. Civit-Masot, L. Miró-Amarante, A. Civit, M. Domínguez-Morales, *Electronics* **2020**, *9*, 1843.
[19] P. Tsvetkova, C. Sousa, D. Beiderbeck, A. M. Kochanowicz, B. Gerazov, M. Agius, T. Przybyła, M. Hoxha, A. H. Tkaczyk, *Disabilities* **2024**, *4*, 1138.

[20] B. Coşkun, P. Uluer, E. Toprak, D. E. Barkana, H. Kose, T. Zorcec, B. Robins, A. Landowska, in *2022 9th IEEE RAS/EMBS Inter. Conf. for Biomedical Robotics and Biomechatronics (BioRob)*, Seoul, Korea, IEEE **2022**, pp. 01–07.

[21] T. L. Praveena, N. M. Lakshmi, in *2021 Third Inter. Conf. on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, IEEE **2021**, pp. 1018–1022.

[22] M. Cheng, Y. Zhang, Y. Xie, Y. Pan, X. Li, W. Liu, C. Yu, D. Zhang, Y. Xing, X. Huang, F. Wang, *IEEE Trans. Affect. Comput.* **2023**, *14*, 2982.

[23] J. Narain, K. T. Johnson, T. F. Quatieri, R. W. Picard, P. Maes, *IEEE Trans. Affect. Comput.* **2022**, *13*, 2238.

[24] O. Banos, Z. Comas-González, J. Medina, A. Polo-Rodríguez, D. Gil, J. Peral, S. Amador, C. Villalonga, *Int. J. Med. Inform.* **2024**, *187*, 105469.

[25] J. L. López-Hernández, I. González-Carrasco, J. L. López-Cuadrado, B. Ruiz-Mezcua, *Sensors* **2019**, *19*, 2620.

[26] J. M. M. Torres, S. Medina-DeVilliers, T. Clarkson, M. D. Lerner, G. Riccardi, *Artif. Intell. Med.* **2023**, *143*, 102545.

[27] A. Sousa, K. Young, M. D'aquin, M. Zarrouk, J. Holloway, *International Conference On Human-Computer Interaction*, Copenhagen, Denmark, Springer **2023**, pp. 657–677.

[28] K. T. Johnson, J. Narain, T. Quatieri, P. Maes, R. W. Picard, *Scientific Data* **2023**, *10*, 523.

[29] C.-M. Wu, S.-C. Chen, Y.-J. Chen, A. Efendi, *IEEE Access* **2024**, *12*, 27115.

[30] C. Regnard, J. Reynolds, B. Watson, D. Matthews, L. Gibson, C. Clarke, *J.Intellect. Disabil. Res.* **2007**, *51*, 277.

[31] E. Ross, C. Oliver, *Br. J. Clin. Psychol.* **2003**, *42*, 81.

[32] M. Roemer, E. Verheul, F. Velthausz, *J. Appl. Res. Intellect. Disabil.* **2018**, *31*, 820.

[33] C. Rowland, *Commun. Disord. Q.* **2011**, *32*, 190.

[34] M. Engelhardt, T. Krämer, M. Marzini, T. Sansour, P. Zentel, *Psychoeducat. Assessment, Intervention Rehabilitation* **2020**, *2*, 1.

[35] T. Hammann, J. Valič, G. Slapničar, M. Luštrek, *Int. J. Dev. Disabil.* **2022**, *70*, 1.

[36] G. Slapničar, E. Dovgan, P. Čuk, M. Luštrek, in *Proc. of the IEEE/CVF Inter. Conf. on Computer Vision (ICCV) Workshops*, Seoul, Korea **2019**.

[37] N. Dalal, B. Triggs, In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. *1*, San Diego, CA, USA, IEEE, **2005**, 886–893.

[38] F. Schroff, D. Kalenichenko, J. Philbin, in *Proc. of the IEEE conference on computer vision and pattern recognition* **2015**, Boston, MA, USA, IEEE, **2015**, pp. 815–823.

[39] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, In *Proc. of the IEEE conference on computer vision and pattern recognition* 2017, Honolulu, HI, USA, IEEE, **2017**, pp. 7291–7299.

[40] B. Přibyl, P. Zemčík, M. Čadík, *Comput. Vis. Image Underst.* **2017**, *161*, 130.

[41] W. Jęśko, *Proc. Interspeech 2021* **2021**, 2921.

[42] S. B. Shah, K. T. Johnson, in *ICASSP 2025-2025 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, IEEE **2025**, pp. 1–5.

[43] J. Lázaro, E. Gil, R. Bailón, A. Mincholé, P. Laguna, *Med. Boil. Eng. Comput.* **2013**, *51*, 233.

[44] N. F. Noy, M. Crubézy, R. W. Fergerson, H. Knublauch, S. W. Tu, J. Vendetti, M. A. Musen, In *AMIA... annual symposium proceedings.*, Washington, DC, USA, AMIA Symp. **2003**, pp. 953–953.

[45] J. Redmon, A. Farhadi, *arXiv* **2018**.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321.

[47] R. H. Horner, E. G. Carr, J. Halle, G. McGee, S. Odom, M. Wolery, *Except. Child.* **2005**, *71*, 165.

[48] B. Maes, S. Nijs, S. Vandesande, I. Van Keer, M. Arthur-Kelly, J. Dind, J. Goldbart, G. Petitpierre, A. Van der Putten, *J. Appl. Res. Intellect. Disabil.* **2021**, *34*, 250.

[49] T. Krämer, P. Zentel, *Psychoeduc. Assess. Intervention Rehabilitation* **2020**, *2*, 15.

[50] T. Karita, Y. Furukawa, Y. Wada, Y. Yagi, S. Senba, E. Onishi, T. Saeki, *JMIR rehabil. assist. technol.* **2021**, *8*, e28020.

[51] V. R. D. M. Herbuela, T. Karita, Y. Furukawa, Y. Wada, A. Toya, S. Senba, E. Onishi, T. Saeki, *Plos one* **2022**, *17*, e0269472.

[52] R. Van Delden, S. Wintels, W. Van Oorsouw, V. Evers, P. Embregts, D. Heylen, D. Reidsma, *J. Intellect. Dev. Disabil.* **2020**, *45*, 66.

[53] G. E. Lancioni, N. N. Singh, M. F. O'Reilly, J. Sigafoos, G. Alberti, V. Perilli, C. Zimbaro, A. Boccasini, C. Mazzola, R. Russo, *J. Intellect. Disabil.* **2018**, *22*, 113.

[54] G. E. Lancioni, N. N. Singh, M. F. O'Reilly, J. Sigafoos, G. Alberti, V. Chiariello, L. Desideri, *Technol. Disabil.* **2021**, *33*, 229.

[55] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, San Francisco, CA, USA, IEEE **2010**, pp. 94–101.

[56] T. Benromano, C. G. Pick, J. Merick, R. Defrin, *Pain Medicine* **2017**, *18*, 441.

[57] B. Simmons, D. Watson, *The PMLD Ambiguity: Articulating The Life-Worlds Of Children With Profound And Multiple Learning Disabilities*, UK, Routledge **2018**.