Transformers and Human-robot Interaction for Delirium Detection

Joe Jeffcock
joejeffcock@gmail.com
Bristol Robotics Laboratory,
University of the West of England
Bristol, U.K.

Mark Hansen*
mark.hansen@uwe.ac.uk
Bristol Robotics Laboratory,
University of the West of England
Bristol, U.K.

Virginia Ruiz Garate
virginia.ruizgarate@uwe.ac.uk
Bristol Robotics Laboratory,
University of the West of England
Bristol, U.K.

ABSTRACT

An estimated 20% of patients admitted to hospital wards are affected by delirium. Early detection is recommended to treat underlying causes of delirium, however workforce strain in general wards often causes it to remain undetected. This work proposes a robotic implementation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) to aid early detection of delirium. Interactive features of the assessment are performed by Human-robot Interaction while a Transformer-based deep learning model predicts the Richmond Agitation Sedation Scale (RASS) level of the patient from image sequences; thermal imaging is used to maintain patient anonymity. A user study involving 18 participants role-playing each of alert, agitated, and sedated levels of the RASS is performed to test the HRI components and collect a dataset for deep learning. The HRI system achieved accuracies of 1.0 and 0.833 for the inattention and disorganised thinking features of the CAM-ICU, respectively, while the trained action recognition model achieved a mean accuracy of 0.852 on the classification of RASS levels during cross-validation. The three features represent a complete set of capabilities for automated delirium detection using the CAM-ICU, and the results demonstrate the feasibility of real-world deployment in hospital general wards.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computer systems organization** → **Robotics**; • **Applied computing** → *Health care information systems*; • **Computing methodologies** → **Vision for robotics**.

KEYWORDS

delirium, transformers, machine learning, machine vision, healthcare

ACM Reference Format:
Joe Jeffcock, Mark Hansen, and Virginia Ruiz Garate. 2023. Transformers and Human-robot Interaction for Delirium Detection . In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3568162.3576971

HRI '23, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9964-7/23/03…$15.00
https://doi.org/10.1145/3568162.3576971

1 INTRODUCTION

Delirium is the onset of confusion resulting from an underlying physical illness and affecting 20% of patients admitted to hospital wards [1]. The condition is associated with poorer clinical outcomes in patients, including higher 6-month mortality rates and extended length of stay in hospitals [13]. It also negatively affects healthcare professionals, where aggressive patient behaviours expose staff to safety risks, increased anxiety, and disruption to medical duties [2]. As a result, significant material and personnel costs are incurred. For example, the average German hospital ward spends 948,000€ annually on delirium patients [37], while the total economic cost of delirium in Australia was estimated at £4.3 billion between 2016 and 2017 [26]. Furthermore, the transfer of susceptible patients from intensive care unit (ICU) to general wards often leads to insufficient care and reporting of delirium, given high patient to nurse ratios [15]. Unforeseen spikes in hospital demand - such as epidemics - are also expected to hinder detection of patient deterioration due to workforce strain [38].

The prevention of delirium directly improves the well-being of patients, healthcare professionals and families by reducing negative physical and mental effects of the condition across all parties [1, 2]. The prevention of long-term complications brought about by delirium, such as dementia [26], present further positive outlook for patients discharged from hospitals. Still, though early detection and prevention is recommended to improve outcomes of delirium, hypoactive symptoms are often misdiagnosed as depression [2].

The mentioned adverse effects of late detection motivate the need for automated assessment of delirium, where human-robot interaction and action recognition technologies can aid healthcare staff in the early detection of symptoms. However, the nature of delirium poses significant research challenges in action recognition, as symptoms present themselves across movement, communication, and thought patterns [1], where the reliance on visual data only limits modes of detection.

Commercially-available action recognition systems often feature low-dimensional thermal infrared input, limiting their applications to coarse tasks such as the detection of falls and movement patterns across rooms. Such systems are not applicable to the task of delirium detection, where fine movements are monitored to test for agitation and sedation [17, 29]. Similar works detect delirium via electroencephalogram (EEG) readings [32, 35]. Yet, the addition of contact sensors can cause discomfort and increase risk of patient harm, while nurses express a need for wireless sensors due to high numbers of cables present at ICU beds [27].

Of the works that utilise machine vision for patient monitoring, many use RGB camera images that can uniquely identify patients [10, 22, 39], negatively impacting data privacy and patient anonymity. Lyra et al. [21] achieve patient monitoring by applying deep learning to thermal images; however, hardware costs limit the scalability of fixed thermal cameras across multiple hospital beds.

Transformers represent a recent shift towards attention mechanisms in deep learning architectures for sequence modelling, outperforming well-established Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) methods while significantly reducing computational complexity by a highly parallelised architecture [36]. Recent advances in transformers for action recognition [23] motivate the feasibility of their use for delirium detection in hospital wards, addressing the limitations of commercially-available systems and related works. Transformers are at the state of the art in natural language processing at the time of writing, and are starting to be applied to machine vision tasks [11, 23] in place of Convolutional Neural Networks (CNNs) to achieve high accuracy at much lower computational cost. CNN, RNN, and LSTM-based methods are shown to be effective at recognising patient activities, however no works were found that explore transformers for vision-based action recognition in medical settings.

Both the CAM-ICU and RASS rely on interaction to determine states of confusion and reactivity to stimuli in patients. Based on the success of previous works in HRI that use multi-modal interaction to enable social tasks [16, 31, 34], this work aims to reduce workforce strain by automating interactive elements of the CAM-ICU traditionally performed by healthcare staff. Speech, haptic, and vision capabilities of humanoid robots [28] identified them as suitable candidates for HRI implementation, while the physical embodiment of interaction creates audio and visual stimuli from which patient attention can be monitored for RASS classification.

We propose an automated delirium detection system based on Machine Vision and Human-robot Interaction (HRI) to help reduce the load of continuous assessment required by trained healthcare professionals. The system implements features of the Confusion Assessment Method for Intensive Care Units (CAM-ICU) used to detect delirium [14] as a robotic system. A socially assistive robot is programmed to detect inattention and disorganised thinking by interviewing patients through HRI, while a transformer-based deep learning model predicts agitation-sedation levels using 2D skeleton keypoints extracted from thermal image sequences of the interaction. The features detected during HRI can then be used to track delirium over time following the CAM-ICU medical algorithm. Delirium detection using thermal cameras thus addresses limitations of low-dimensional and contact-based action recognition in healthcare, while thermal imaging and 2D skeleton extraction increases anonymity by reducing the number of features that can be used to uniquely identify patients.

2 METHODOLOGY

2.1 Overview

This project proposes an adaptation of the CAM-ICU algorithm for delirium detection [14] into a robotic system to aid healthcare staff in wards with high patient-nurse ratios. The method involves the diagnosis of four features for use in a medical algorithm (Figure 1) to determine if a patient is delirium positive or negative. Of the four features, two are based on interaction with healthcare staff, while one measures the state of the patient using other tools such as the RASS, described in further detail below. The last features tracks the fluctuation of state over time, accounting for the sudden onset of symptoms associated with delirium. The CAM-ICU was validated in a study across 111 mechanically ventilated patients [12].

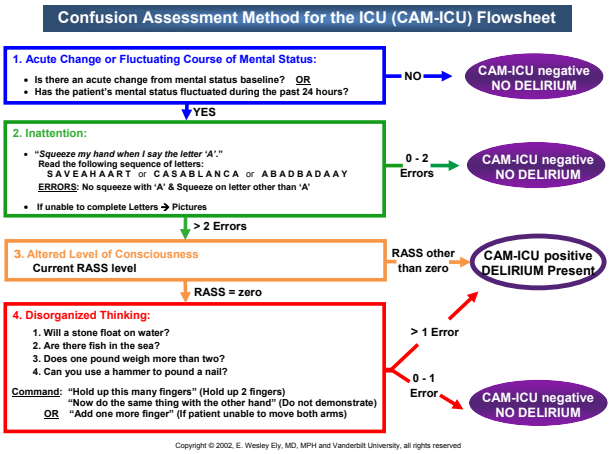


Figure 1: Flowsheet of the CAM-ICU algorithm. Excerpt from the CAM-ICU educational materials (unrestricted in terms of use).

The Richmond Agitation-Sedation Scale (RASS) is a 10-level scale describing the level of sedation or agitation of a patient, aimed at assessing the appropriate dosage of sedative medication administered to patients in ICU wards [29]. The scale ranges from "unarousable" (-5) to "alert" (0) to "combative" (+4), specifying a 3-step procedure to test for features of each level. The RASS achieved high inter-rater reliability among five investigators assessing sedation in its validation study in the ICU [29], and is recommended across two features in the CAM-ICU to detect delirium.

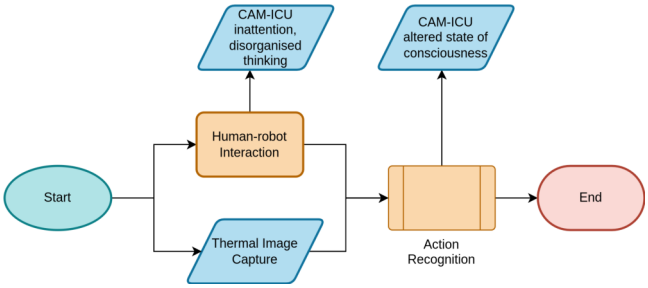
Of the four features, Features 2 and 4 - Inattention and Disorganised Thinking - were deemed appropriate for human-robot interaction (HRI) solutions, requiring logic across behaviours involving multi-modal interaction. The SoftBank Robotics Pepper platform was chosen for HRI, as it features off-the-shelf access to capabilities required by the CAM-ICU, such as touch sensing, text-to-speech, speech recognition, and camera input [28].

Motivated by related work in patient monitoring by [32, 39], this work aims to classify Feature 3 - Altered level of conciousness - by predicting RASS scores through a machine learning model trained on time-series data. Differing from the use of EEG signals to predict sedation-agitation levels in [32], this work proposes machine vision-based RASS classification using the AcT transformer model, taking time-series, 2D skeleton data extracted from a thermal image device using OpenPose as an input.

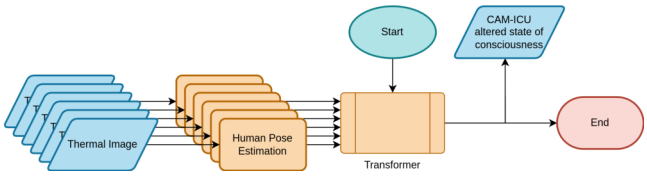
Feature 1 - Acute Onset - tracks patient status over time, and thus the implementation of Features 2-4 described above provide full coverage of features required to detect delirium in hospitals

using CAM-ICU. The following subsections outline the research methodology of developing a robotic system for delirium detection, including the implementation, testing, and validation of the proposed modules.

The initial concept design of the proposed system is depicted in Figure 2. Skeleton keypoints from human pose estimation were chosen as visual features to be extracted from thermal images for the action recognition model, due to their wide success across previous works [3, 19, 23, 24, 30, 40].



(a) CAM-ICU feature detection



(b) Action Recognition sub-process

Figure 2: System Concept Design

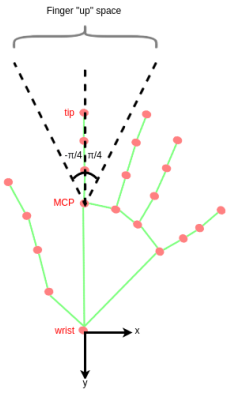
Sliding windows and de-noising are implemented to classify frame sequences of arbitrary length using the fixed input size AcT architecture, described in further detail in Section 2.2.

2.2 Implementation

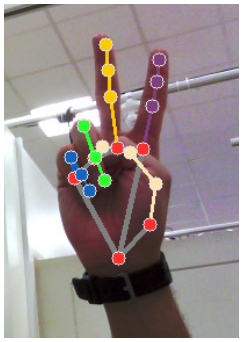
CAM-ICU Feature 2: letters attention via touch sensing. In detecting the inattention feature of the CAM-ICU, the robot utters each letter of the word "SAVEAHAART" while users are required to tap the top of its hand upon hearing the letter "A". A beep noise is emitted on touch input as audio feedback. User input is compared against the ground truth to compute the absolute error.

CAM-ICU Feature 4 (1): yes/no questions and speech recognition. The first component of the disorganised thinking feature is detected by speech interaction based on four yes/no logic questions asked by the robot: "can a stone float on water?", "are there fish in the sea?", "does one pound weigh more than two pounds?", "can you hit a nail with a hammer?". Speech recognition will register user input for each question, compared against ground truth to compute the absolute error.

CAM-ICU Feature 4 (2): gesture recognition for commands. The second component of the disorganised thinking feature is detected by hand gesture recognition based on a command given by the robot. Hand gesture recognition is built on Google's MediaPipe Hands API [20] and implemented as a TCP/IP server to accept requests from NAOqi scripts. For each finger in a hand detected from an image using MediaPipe, the angle from the wrist to the metacarpophalangeal joint (MCP) is aligned along the y-axis of the image frame such that the finger - defined from the MCP to the tip - is assumed to be "up" at an angle between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$, and "down" otherwise. The result of gesture recognition is a 5-bit string indicating "up" or "down" for each finger, along with a string indicating the handedness ("Left", "Right") of the detection.



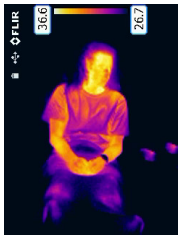
(a)



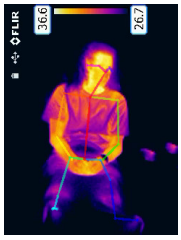
(b)

Figure 3: MediaPipe for hand state prediction. (a) Visual depiction of finger state prediction in the image x, y frame. (B) Example of a hand state prediction "01100 Right" (note that input images are assumed to be mirrored)

The detection process is as follows. The robot first displays an image of two fingers and asks the user to hold up as many fingers. The most common detected hand gesture out of 10 is then taken as a user response, represented by the process "hand state prediction". Following this, the robot will ask the user to repeat the process with their other hand. The error defaults to 0, where any incorrect user response results in an error of 1.



(a) Input image



(b) OpenPose skeleton overlay

Figure 4: 2D skeleton extraction on thermal images using OpenPose

CAM-ICU Feature 3: image-based RASS prediction. The OpenPose human pose estimation library was used to achieve skeleton keypoint extraction in this project, packaged in a CUDA-enabled Docker container for GPU support and portability of code. The

package performs a forward pass of OpenPose on each frame of each input video, and outputs the extracted skeleton keypoints and labels for each frame. Figure 4 shows an example of 2D skeleton extraction on an input thermal image.

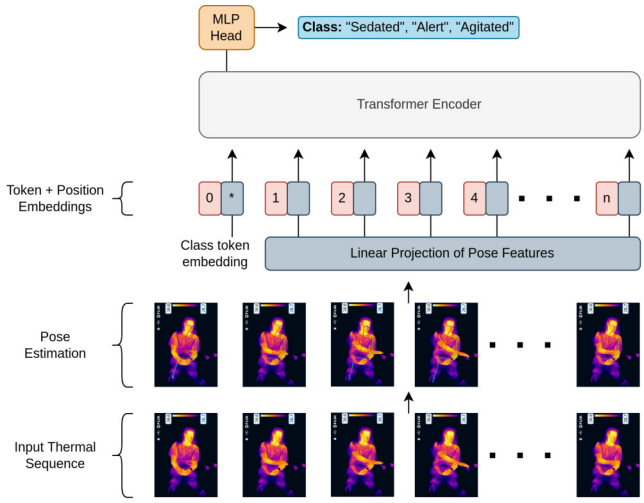


Figure 5: Overview of the AcT architecture applied to thermal image-based RASS classification. Figure adapted from the AcT architecture overview in [23].

As in skeleton keypoint extraction, the AcT transformer model is implemented in a Docker package for portability of code. The original implementations of AcT and the MPOSE2021 data loader were adapted to support custom datasets and use in other applications. Figure 5 depicts an overview of the proposed architecture.

Given the task of classifying video input of arbitrary length, a sliding window parameterised by frameskip, size, and stride was implemented to extract sequences of frames for use in the fixed input size AcT architecture. Frameskip determines the sampling rate of frames from input data, where higher values extract data across a longer duration at the cost of losing finer motion information captured in skipped frames. Size and stride control the number of frames included in the window and its step size as it slides across time series data, respectively.

The use of sliding windows to extract subsets of input data can result in noisy predictions where different frames of the same video are given different classifications. This work proposes to de-noise outputs by taking the mode of predictions across a video as its classification; this process is visible in Figures 9 and 10.

2.3 Action Recognition Training Methodology

Based on the CAM-ICU flow sheet in Figure 1, the task of action recognition for detecting an altered state of consciousness is to predict the RASS level of a user given a video of their interaction with the HRI component of the proposed solution. A RASS dataset was collected for this task during the user study described in Section 2.4, comprising videos of participants completing the HRI task portraying one of each of the alert, sedated and agitated RASS states. For each video — organised by actor — data points are generated from each frame by extracting the 2D skeleton of the participant using OpenPose and labelling it with the RASS level portrayed.

Due to the small sample size of participants in the dataset and high similarity within data points of the same actor in successive video frames, 6-fold cross-validation between actors is proposed to assess the performance and generalisability of machine learning solutions to RASS prediction. For each of the 6 folds, the dataset is thus split into 15 actors for training with 3 actors held out as unseen data for the test set. Data points in the training set - regardless of actor - are further split randomly into training and validation subsets in a 10 : 1 ratio to validate performance at train time.

The Support Vector Machine (SVM) was chosen as a baseline for the RASS task, as it directly uses features in the data to find hyperplanes for classification [9], opposed to learned attention in AcT. The same 30-frame sequences of 2D skeleton features were used across both methods, albeit flattened in the SVM since it does not incorporate positional embeddings present in transformers.

AcT was also trained on the benchmark MPOSE2021 action recognition dataset [23] for the comparison of metrics against the RASS task and for use as a base model for transfer learning.

Training		Regularisation	
Training epochs	350	Weight decay	1e-4
Batch size	512	Label smoothing	0.1
Optimizer	AdamW	Dropout	0.3
Warmup epochs	30%	Random flip	0%
Step epochs	80%	Random noise σ	0.0

Table 1: AcT model training and regularisation parameters

Based on empirical hyperparameter tuning, the following parameters for AcT and the SVM that obtained the best mean balanced accuracy during grid search were selected to train models for the final results in Section 3: **AcT** `{"frameskip":30, "size":30, "stride":1}`; **SVM** `{"kernel":"RBF", "C":10, "gamma":0.001}`.

All other training and regularisation parameters of AcT are set to the values in Table 1, while all other parameters of the SVM are set to the scikit-learn [25] defaults.

2.4 Experimental set-up

The set-up for the user study is depicted in Figure 6a. A hospital-like environment is recreated with an adjustable bed and blanket, where participants are instructed to remove any footwear and sit as shown in the diagram. The Pepper robot is positioned diagonally to the left of the user and programmed to detect Features 2 (Inattention) and 4 (Disorganised Thinking) of the CAM-ICU in sequence, following the implementation described in Section 2.2.

A 160×120px resolution FLIR E30 thermal imaging camera is mounted on a tripod 2.25m in front of the user at a height of 1.6m, oriented in portrait such that the entirety of the bed is captured in frame (Figure 6b) - measurements are subject to change depending on the field of view of the camera lens. The temperature range for thermal imaging is set to 27.5–37.6 degrees Celcius to capture human surface temperatures as well as heat transfer to clothing

(a) (b)

Figure 6: (a) Panoramic view of the user study set-up (b) Fixed camera set up in the user study

and bedding. Optionally, an RGB camera is placed next to the thermal camera to mitigate project risks related to human keypoint extraction in thermal images.

Due to the lack of a publicly available dataset for thermal video-based delirium detection, the user study involved 18 participants comprised of researchers at the Bristol Robotics Lab, selected in line with the Ethical Review Checklist for Postgraduate Taught Modules. All participants were assumed to be in an alert state (RASS = 0). Participants were briefed on the purpose and procedure of the user study, as well as on agitated, sedated, and alert behaviour patterns present in the RASS.Before the actual experimental recording, participants underwent the HRI-based CAM-ICU once to be familiarised with the system. Then, each each participant underwent the HRI-based CAM-ICU under 3 conditions in a random order: in an alert state with scores collected as quantitative results, role-playing as a patient with RASS > 0 (agitated), and role-playing as a patient with RASS < 0 (sedated).

A total of 54 labelled videos were collected from 18 participants in the user study, each having completed or role-played the task in the alert, sedated, and agitated states. The three states correspond to zero, negative, and positive scores on the RASS, respectively. The data comprises 244415 frames over 8147.38 seconds (135 minutes, 47 seconds), each depicting a thermal image containing a participant in the fixed camera view as shown in Figure 6b.

2.5 Data Collection and Metrics

The following data was collected for each of the 18 participants: personal details (occupation, field, age, gender, height, weight, first language); RGB and thermal video data of engagement in HRI in three states — alert with RASS = 0, role-playing RASS > 0, and role-playing RASS < 0. The following dependent measures were collected during the experiments:

- Inattention errors from 0-10 recorded in the alert assessment.
- Disorganised thinking (1) errors from 0-4 recorded in the alert assessment.
- Disorganised thinking (2) errors from 0-1 recorded in the alert assessment.
- Time taken for each HRI task.

The following metrics were used to measure the success of the HRI, person detection, and action recognition modules:

- **Absolute error:** Absolute error ϵ_{abs} is defined as the total number of incorrect values output by a system, yielding an absolute measure of the total number of failure cases.
- **Relative error:** Relative error $\epsilon_{rel} = \frac{\epsilon_{abs}}{N}$ is defined as a fraction of the absolute error over the total number of test cases N. The metric describes the failure of a module over the total number of test cases.
- **Accuracy:** Accuracy a is defined as the total number of correct values output by a system over the total number of test cases N. The metric describes the success of a module over the total number of test cases.

Additionally, **Balanced accuracy** was selected to measure the success of the action recognition component, based on metrics presented in [23]. Balanced accuracy is defined as the average of recall calculated for each class [4, 25], describing the success of the model accounting for imbalanced datasets.

3 RESULTS

A total of 54 labelled videos were collected from 18 participants in the user study, each having completed or role-played the task in the alert, sedated, and agitated states. The three states correspond to zero, negative, and positive scores on the RASS, respectively. The data comprises 244415 frames over 8147.38 seconds (135 minutes, 47 seconds), each depicting a thermal image containing a participant in the fixed camera view as shown in Figure 6b.

3.1 HRI Results

Feature	Abs. error	Rel. error	False positives	Acc.
Inattention	0	0.0	0	1.0
(1)	10	0.139	3	0.833
(2)	6	0.333		

Table 2: Summary of participant performance in the HRI user study. (1) and (2) refer to Disorganized Thinking

From Table 2, No errors were recorded from participants under assessment for the inattention feature of the CAM-ICU, yielding no false positives (FP). This results in an accuracy of 1.0 for Feature 2 of the CAM-ICU.

Table 2 further shows that 10 errors were encountered in the logical question part (DT1) of the disorganised thinking feature, along with 6 errors in the gesture command part (DT2). Four of the errors in DT1 resulted from non-detection in the speech recognition engine, while the remaining six were due to incorrect participant answers. As for DT2 errors, three resulted from non-detection of hands by MediaPipe, with the remaining three caused by the miscounting of fingers in participant hand gestures.

Three false positives were identified in the HRI user study, all of which occurring in the disorganised thinking feature (Table 2). Two of the false positives exhibit incorrect responses in DT1 and non-detection in DT2, while the remaining one results from non-detection in both DT1 and DT2. The Inattention module had the best performance with a relative error of 0.0, while the DT2 module performed the worst with a relative error of approximately 0.333.

The mean and standard deviation of the time taken to complete the HRI task were computed to be approximately 143.88s and 8.79s, respectively, while the maximum and minimum duration recorded were 158.33s and 117.27s, respectively.

3.2 Person Detection Results

It can be assumed that exactly one human participant is captured in every labelled thermal image frame obtained from the user study. As such, metrics in Section 2.5 are computed for each frame by the number of humans extracted by OpenPose, given a target value of exactly 1.

Metric	Cumulative	Per Participant
Total frames	244415	-
Absolute error	3692	-
Relative error	0.015	0.014 ± 0.034
Accuracy	0.985	0.986 ± 0.034

Table 3: Metrics computed across the entire user study dataset

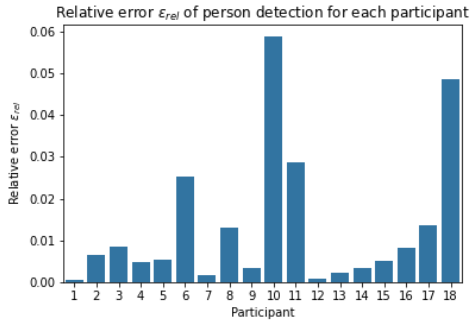


Figure 7: Relative error of person detection

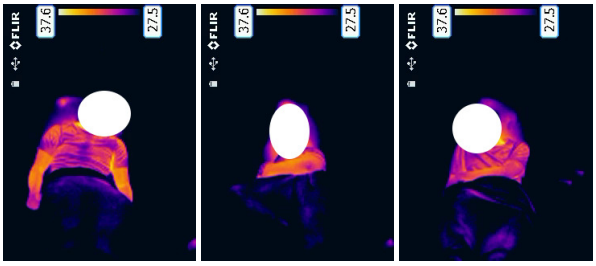


Figure 8: Examples of frames where no persons were detected; participants are anonymised for privacy.

Table 3 tabulates person detection metrics computed on the user study dataset. From Figure 7, it can be seen that relative error is not evenly distributed among participants, where participants 10 and 18 show significantly higher relative errors than others in the dataset. Qualitatively, Figure 8 shows three images from these participants where no persons were detected. Participants are seen to assume slouched postures, where generalisation to thermal images may prove challenging due to a lack of representative data in the RGB dataset on which OpenPose was trained.

3.3 Action Recognition Results

	Train	Validation	Test		
Fold	Accuracy	Accuracy	Acc.	Bal.	Mode
1	0.999	1.0	0.776	0.777	0.889
2	0.999	0.996	0.716	0.717	0.889
3	1.0	1.0	0.645	0.653	0.778
4	0.996	1.0	0.770	0.765	0.889
5	0.999	1.0	0.737	0.732	0.889
6	1.0	1.0	0.679	0.68	0.778
Mean	0.999	0.999	0.720	0.721	0.852
SD	0.001	0.001	0.047	0.044	0.052

Table 4: Metrics obtained during 6-fold cross-validation of AcT on the thermal RASS dataset

Table 4 tabulates the final train, validation and test set metrics obtained from AcT during 6-fold cross-validation on the RASS prediction task (Section 2.4) using hyperparameters in Section 2.3.

The generated loss curves were consistent over all folds and converged by a maximum of 82 epochs using early stopping (patience = 15) and the model that obtained the highest validation accuracy was selected for testing at the end of each fold.

From Table 4, it is seen that AcT attained high train and validation accuracies across all folds; both sets showing a mean accuracy of 0.999 and standard deviation of 0.001. The train and validation accuracies are significantly higher than that of the test set, with a mean of 0.720 and standard deviation of 0.047. Given that each class appears in an equal number of videos in the RASS dataset (Section 2.4), AcT is seen to obtain similar accuracy and balanced accuracy scores across all folds in the RASS prediction task. De-noising by taking the mode of each video is shown to exhibit increased performance over frame-level prediction, where mode accuracy is greater the accuracy and balanced accuracy metrics computed for all folds in cross-validation (Figure 4).

Table 5 tabulates scoring metrics on the RASS task, organised by actor. Comparing Tables 4 and 5, mean balanced accuracy is similar when organising results by fold and actor, at 0.721 and 0.714, respectively. However, it can be seen that standard deviation is significantly higher between actors over folds, recorded at 0.155 and 0.044, respectively. The highest mean balanced accuracy was achieved on actor 11 in fold 1 (0.919), while the lowest was in actor 7 of fold 5 (0.43). No mode accuracy below 0.666 was recorded for any actor, while AcT achieved mode accuracy of 1.0 for 10 actors during cross-validation.

Qualitatively, Figures 9 and 10 visualise the performance of AcT, before and after de-noising, respectively. Each fold comprises 3 actors portraying 1 of each of the alert, sedated and agitated RASS states. Videos are delimited by transitions in the target column value for each actor. As an example, the target column of actor 6

Fold	Actor	Accuracy	Balanced	Mode
1	1	0.894	0.893	1.0
	11	0.925	0.919	1.0
	12	0.482	0.484	0.667
2	0	0.595	0.597	0.667
	5	0.755	0.755	1.0
	14	0.805	0.799	1.0
3	6	0.44	0.44	0.667
	8	0.648	0.696	0.667
	13	0.821	0.814	1.0
4	4	0.826	0.832	1.0
	9	0.628	0.615	0.667
	10	0.856	0.84	1.0
5	2	0.857	0.854	1.0
	7	0.48	0.43	0.667
	17	0.827	0.831	1.0
6	3	0.535	0.554	0.667
	15	0.66	0.669	0.667
	16	0.83	0.837	1.0
Mean		0.715	0.714	0.852
SD		0.153	0.155	0.166

Table 5: Test metrics obtained during 6-fold cross-validation of AcT, organised by actor.

contains three sequences of agitated, sedated, and alert frames, in order; each sequence represents a video of the actor portraying the target RASS level. Meanwhile, the predicted column of Figure 9 shows the Fold 2 AcT model exhibiting a number of errors in RASS prediction, with "agitated" frames predicted as "sedated", "sedated" frames predicted as "agitated", and so on. By taking the mode prediction for each video sequence in Figure 9, the de-noised result in 10 is obtained. In the case of actor 6, all errors are filtered by de-noising, leading to the predicted column matching the target column in Figure 10.

The effects of de-noising can be observed across all actors and folds, where inconsistent frame-level predictions in Figure 9 are filtered to 46 correctly classified videos out of 54 in Figure 10.

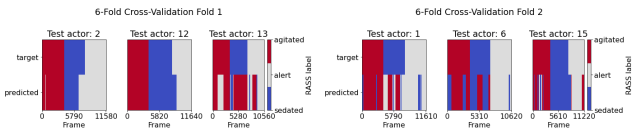


Figure 9: Example visual representation of AcT results on the collected dataset

Table 6 compares the performance of AcT against other baseline models on the RASS task, as well as against AcT trained on the benchmark MPOSE2021 action recognition dataset.

The hyperparameters selected in Section 2.3 were used to train AcT and the SVM on the RASS task, while those in Table 1 were used

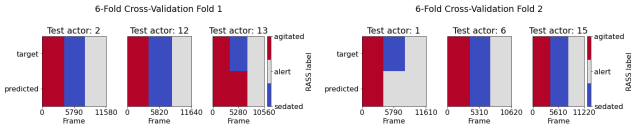


Figure 10: Example visual representation of AcT results on the collected dataset, de-noised.

to train AcT on MPOSE2021. The results of a dummy classifier, set to always predict the most frequent label in the training set given any input, are included in the comparison, demonstrating performance when all input features are ignored during classification.

Model	Task	Transfer learning	Pre-trained on	Mean accuracy	Mean balanced	Mean mode
AcT	RASS	No	-	0.720	0.721	**0.852**
AcT	RASS	Yes	MPOSE'21	0.718	0.718	0.833
SVM	RASS	No	-	0.686	0.686	0.796
Dummy	RASS	No	-	0.333	0.333	0.333
AcT	MPOSE'21	No	-	**0.889**	**0.841**	-

Table 6: Accuracy, balanced accuracy, and mode accuracy metrics obtained across different models and action recognition tasks. The mean mode accuracy of AcT on RASS exceeds the mean balanced accuracy of AcT on MPOSE2021.

From Table 6, all models that leverage features (AcT, SVM) outperformed the dummy classifier baseline which scored 0.333 across all three accuracy metrics, consistent with a near-equal split between alert, agitated, and sedated labels in the RASS dataset. Trained from random initial weights, AcT scored the highest across all three metrics in the RASS task, with a mean accuracy, balanced accuracy, and mode accuracy of 0.720, 0.721, and 0.852, respectively. Training only the final dense layers of AcT, transfer learning from the MPOSE2021 dataset obtained slightly lower results than the model trained fully on the RASS task. The SVM model obtained the lowest scores of the models trained on the RASS task, however significantly outperforming the dummy classifier with mean accuracy, balanced accuracy, and mode accuracy scores of 0.686, 0.686, and 0.796, respectively. In comparing different action recognition tasks, the mean accuracy and balanced accuracy of AcT on MPOSE2021 were significantly higher than those of all models on the RASS task.

4 DISCUSSION

The performance results of the user study in Table 2 show that the HRI implementations of Features 2 and 4 of the CAM-ICU scored accuracies of 1.0 and 0.833, respectively, with three false positives (FP) for Disorganised Thinking (Feature 4) out of 18 participants in total. As already mentioned, these were found to be due to either incorrect responses or their response not being detected. Alternative text-to-speech agents or text output to a screen may be considered to improve the clarity of speech during DT1.

Additionally, our results show that FP participants were the only cases where hand gestures were not detected in DT2. Observations

during the study revealed that gestures were outside the field of view of the Pepper robot's camera in such cases, where detailed usage instructions or a shorter focal length camera lens may reduce such errors. Aside from non-detection, the number of fingers in DT2 were miscounted on three occasions, contributing to a higher relative error of 0.333 in Table 2. Counting errors were often observed as false positives in the thumb or ring finger, indicating that more robust geometry incorporating joints between the finger MCP and tip (Figure 3a) is required for accurate hand state prediction.

RASS prediction for Feature 3 of the CAM-ICU achieved a mean accuracy of 0.852 when taking the mode prediction of each video as its classification (Table 4). While the trained AcT model distinguishes only between positive, zero, and negative RASS scores, the results outperform previous work in delirium detection using deep learning on EEG signals, where [32] shows a median accuracy of 0.7 predicting RASS between 0 and −5 while allowing for 1 level of difference to constitute a true positive. On the other hand, [35] achieves a sensitivity of 1.0 and specificity of 0.96 on delirium detection (see Section 1), however do not include a train-test split to test for generalisation, and are thus comparable to the mean training accuracy of 0.999 of AcT (Table 4) where predictions are made on examples previously seen by the model. The results of RASS prediction carried out in this work are thus shown to be on par with related works in the field of delirium detection, prompting future work into the collection of data to test scalability and generalisability in real-world hospital settings.

From a machine vision perspective, the comparison between the AcT and SVM models on the RASS prediction task highlight the effectiveness of 2D skeleton keypoints as features for action recognition. From Table 6, the SVM is seen to exhibit performance close to that of the AcT models, at a mean balanced accuracy of 0.686 compared to 0.721 recorded for AcT. Given the lack of attention mechanisms for learning patterns in the data, relatively high performance of the SVM model reflects on the need for effective feature extraction in action recognition problems. Table 6 also demonstrates that AcT retains effective transfer learning capabilities present in the Vision Transformer [11] from which its architecture was adapted. This is evidenced by an AcT model, pre-trained on MPOSE2021 and transferred to the RASS task by training only the final two layers, approaching the performance of a similar model trained from scratch on the RASS dataset.

Overall, the proposed system for delirium detection achieved accuracies of 1.0 and 0.833 for Features 2 (inattention) and 4 (disorganised thinking) of the CAM-ICU, respectively, and a mean accuracy of 0.852 on Feature 3 (altered state of consciousness) when the mode RASS prediction is taken for each input video. Results were obtained across a relatively diverse set of participants in terms of gender and body type, and first language (nine different first languages), promoting the generalisability of the machine vision for action recognition and HRI components respectively.

The implementation of Features 2-4 represent a complete solution to the robotic detection of features in the CAM-ICU, where inattention and disorganised thinking are determined by HRI, while an altered state of consciousness is predicted from a video of the interaction by a trained AcT model. When deployed in a hospital setting, Feature 1 (sudden onset) can be computed by tracking fluctuations in Feature 3 over time. The mean time of the assessment was recorded at 143.88s with a standard deviation of 8.79s (Section 3.1), meaning a single robot may assess up to 25 patients per hour. Given comparable performance to previous works in delirium detection using EEG electrodes attached to patients [32, 35], the results of the proposed system show promise towards a contact-free solution to automated delirium detection in hospital wards.

However, a number of limitations were identified in the user study in Section 2.4 affecting the validity and reliability of the HRI results and data collected for the RASS dataset. The external validity of the user study is limited by the use of alert and healthy participants as well as the laboratory-based set up in Figure 6a. By only involving alert participants role-playing different RASS levels, the HRI results and trained models may not reflect performance on the target group of patients in hospital wards, where responses to the robot and RASS behaviours exhibited may differ significantly in real-life settings. Participants were briefed on RASS levels using the CAM-ICU training manual, however information on behaviours may be interpreted differently across participants. As such, portrayals of agitation and sedation in collected data may be inconsistent or inaccurate to actual behaviours in patients with delirium.

While the goal of this work was to evaluate the technical feasibility of a HRI system for autonomous delirium detection, user perception and acceptance are key considerations that will need to be addressed in future real-world implementations. Aspects of this can include the influence of robot anthropomorphism [18], the acceptance of different interfaces and AI by clinicians [5–7], safety implications [8], and gender-sensitive approaches [33].

5 CONCLUSIONS

We have presented the design, development and testing of a robotic implementation of the CAM-ICU for automated detection of delirium in hospital wards. HRI routines and deep learning-based agitation-sedation prediction were implemented to detect features of the CAM-ICU; an 18-participant user study was conducted to validate performance of the former modules while collecting machine learning data for the latter.

The results demonstrate the feasibility of robotic systems and Transformer-based action recognition to aid delirium detection in hospitals facing high workforce strain. Future work is required to scale the user study in Section 2.4 to a real-world hospital ward, aimed at addressing limitations to the external validity and reliability of this work, as well as testing the feasibility of the proposed system for automated delirium detection. Results obtained from a hospital ward would function to validate performance the target group of the system, using reliable data from real-world patients. The challenges of a real-world implementation would also motivate improvements to HRI highlighted in Section 4, while visual representations of attention in AcT - such as plots of attention weights in [23] - can be explored to increase the explainability of action recognition for ethical use in medical settings.

ACKNOWLEDGMENTS

We would like to acknowledge the Occupational Therapist and Improvement Specialist from the Acute NHS trust who kindly provided their time for a discussion on ICU needs.

REFERENCES

[1] Yasmin Ahmed, Phil Timms, and Thomas Kennedy. 2019. *Delirium.* [Online]. [Accessed 21 March 2022] Available from https://www.rcpsych.ac.uk/mental-health/problems-disorders/delirium.

[2] Emma Arend and Martin Christensen. 2009. Delirium in the intensive care unit: a review. *Nursing in Critical Care* 14, 3 (2009), 145–154.

[3] Danilo Avola, Marco Cascio, Luigi Cinque, Gian Luca Foresti, Cristiano Massaroni, and Emanuele Rodolà. 2019. 2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs. *IEEE Transactions on Multimedia* 22, 10 (2019), 2481–2496.

[4] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*. 3121–3124. https://doi.org/10.1109/ICPR.2010.764

[5] Francisco M Calisto, Alfredo Ferreira, Jacinto C Nascimento, and Daniel Gonçalves. 2017. Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. 390–395.

[6] Francisco Maria Calisto, Nuno Nunes, and Jacinto C Nascimento. 2020. BreastScreening: on the use of multi-modality in medical imaging diagnosis. In *Proceedings of the international conference on advanced visual interfaces*. 1–5.

[7] Francisco Maria Calisto, Nuno Nunes, and Jacinto C Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (2022), 102922.

[8] Antonella Camilleri, Sanja Dogramadzi, and Praminda Caleb-Solly. 2022. Learning from Carers to inform the Design of Safe Physically Assistive Robots-Insights from a Focus Group Study. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 703–707.

[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[10] Anis Davoudi, Kumar Rohit Malhotra, Benjamin Shickel, Scott Siegel, Seth Williams, Matthew Ruppert, Emel Bihorac, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Azra Bihorac, et al. 2019. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Scientific reports* 9, 1 (2019), 1–13.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] E. Wesley Ely, Sharon K. Inouye, Gordon R. Bernard, Sharon Gordon, Joseph Francis, Lisa May, Brenda Truman, Theodore Speroff, Shiva Gautam, Richard Margolin, Robert P. Hart, and Robert Dittus. 2001. Delirium in Mechanically Ventilated PatientsValidity and Reliability of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *JAMA* 286, 21 (12 2001), 2703–2710. https://doi.org/10.1001/jama.286.21.2703 arXiv:https://jamanetwork.com/journals/jama/articlepdf/194422/jce10051.pdf

[13] E Wesley Ely, Ayumi Shintani, Brenda Truman, Theodore Speroff, Sharon M Gordon, Frank E Harrell Jr, Sharon K Inouye, Gordon R Bernard, and Robert S Dittus. 2004. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *Jama* 291, 14 (2004), 1753–1762.

[14] Wesley Ely. 2016. *Confusion Assessment Method for the ICU (CAM-ICU) - The Complete Training Manual.* ICU Delirium. [Online]. [Accessed 21 August 2022] Available from https://www.icudelirium.org/medical-professionals/delirium/monitoring-delirium-in-the-icu.

[15] Ronny Enger and Birgitta Andershed. 2018. Nurses' experience of the transfer of ICU patients to general wards: a great responsibility and a huge challenge. *Journal of clinical nursing* 27, 1-2 (2018), e186–e194.

[16] Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald P.A. Petrick. 2012. Two People Walk into a Bar: Dynamic Multi-Party Social Interaction with a Robot Agent *(ICMI '12)*. Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/2388676.2388680

[17] Sharon K Inouye, Christopher H van Dyck, Cathy A Alessi, Sharyl Balkin, Alan P Siegal, and Ralph I Horwitz. 1990. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Annals of internal medicine* 113, 12 (1990), 941–948.

[18] Kim Klüber and Linda Onnasch. 2022. Appearance is not everything-Preferred feature combinations for care robots. *Computers in Human Behavior* 128 (2022), 107128.

[19] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*. Springer, 816–833.

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. https://doi.org/10.48550/ARXIV.1906.08172

[21] Simon Lyra, Leon Mayer, Liyang Ou, David Chen, Paddy Timms, Andrew Tay, Peter Y Chan, Bergita Ganse, Steffen Leonhardt, and Christoph Hoog Antink. 2021. A deep learning-based camera approach for vital sign monitoring using thermography images for ICU patients. *Sensors* 21, 4 (2021), 1495.

[22] Andy J Ma, Nishi Rawat, Austin Reiter, Christine Shrock, Andong Zhan, Alex Stone, Anahita Rabiee, Stephanie Griffin, Dale M Needham, and Suchi Saria. 2017. Measuring patient mobility in the ICU using a novel noninvasive sensor. *Critical care medicine* 45, 4 (2017), 630.

[23] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. 2022. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition* 124 (4 2022), 108487. https://doi.org/10.1016/j.patcog.2021.108487

[24] Farzan Majeed Noori, Benedikte Wallace, Md Uddin, Jim Torresen, et al. 2019. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Scandinavian conference on image analysis*. Springer, 299–310.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Lynne Pezzullo, Jared Streatfeild, Josiah Hickson, Andrew Teodorczuk, Meera R Agar, and Gideon A Caplan. 2019. Economic impact of delirium in Australia: a cost of illness study. *BMJ open* 9, 9 (2019), e027514.

[27] Akira-Sebastian Poncette, Lina Mosch, Claudia Spies, Malte Schmieding, Fridtjof Schiefenhövel, Henning Krampe, Felix Balzer, et al. 2020. Improvements in patient monitoring in the intensive care unit: survey study. *Journal of medical Internet research* 22, 6 (2020), e19091.

[28] SoftBank Robotics. 2022. *Pepper (NAOqi 2.5).* SoftBank Robotics. [Online]. [Accessed 10 September 2022] Available from https://developer.softbankrobotics.com/pepper-naoqi-25.

[29] Curtis N Sessler, Mark S Gosnell, Mary Jo Grap, Gretchen M Brophy, Pam V O'Neal, Kimberly A Keane, Eljim P Tesoro, and RK Elswick. 2002. The Richmond Agitation–Sedation Scale: validity and reliability in adult intensive care unit patients. *American journal of respiratory and critical care medicine* 166, 10 (2002), 1338–1344.

[30] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[31] Rainer Stiefelhagen, Hazim Kemal Ekenel, Christian Fugen, Petra Gieselmann, Hartwig Holzapfel, Florian Kraft, Kai Nickel, Michael Voit, and Alex Waibel. 2007. Enabling Multimodal Human–Robot Interaction for the Karlsruhe Humanoid Robot. *IEEE Transactions on Robotics* 23, 5 (2007), 840–851. https://doi.org/10.1109/TRO.2007.907484

[32] Haoqi Sun, Eyal Kimchi, Oluwaseun Akeju, Sunil B Nagaraj, Lauren M McClain, David W Zhou, Emily Boyle, Wei-Long Zheng, Wendong Ge, and M Brandon Westover. 2019. Automated tracking of level of consciousness and delirium in critical illness using deep learning. *NPJ digital medicine* 2, 1 (2019), 1–8.

[33] Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson, and Ginevra Castellano. 2022. Gender Fairness in Social Robotics: Exploring a Future Care of Peripartum Depression.. In *HRI*. 598–607.

[34] Panagiota Tsarouchi, Sotiris Makris, and George Chryssolouris. 2016. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing* 29, 8 (2016), 916–931. https://doi.org/10.1080/0951192X.2015.1130251 arXiv:https://doi.org/10.1080/0951192X.2015.1130251

[35] Arendina W. van der Kooi, Irene J. Zaal, Francina A. Klijn, Huiberdina L. Koek, Ronald C. Meijer, Frans S. Leijten, and Arjen J. Slooter. 2015. Delirium Detection Using EEG. *Chest* 147, 1 (2015), 94–101. https://doi.org/10.1378/chest.13-3050

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[37] W Weinrebe, E Johannsdottir, M Karaman, and I Füsgen. 2016. What does delirium cost? *Zeitschrift für Gerontologie und Geriatrie* 49, 1 (2016), 52–58.

[38] John Willan, Andrew John King, Katie Jeffery, and Nicola Bienz. 2020. Challenges for NHS hospitals during covid-19 epidemic.

[39] Serena Yeung, Francesca Rinaldo, Jeffrey Jopling, Bingbin Liu, Rishab Mehra, N Lance Downing, Michelle Guo, Gabriel M Bianconi, Alexandre Alahi, Julia Lee, et al. 2019. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *NPJ digital medicine* 2, 1 (2019), 1–5.

[40] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*. 2117–2126.