IEEE Journal of Translational Engineering in Health and Medicine

Received 27 September 2023; revised 4 March 2024 and 22 March 2024; accepted 2 May 2024.
Date of publication 7 May 2024; date of current version 24 May 2024.

Digital Object Identifier 10.1109/JTEHM.2024.3397737

A Dual-Camera Eye-Tracking Platform for Rapid Real-Time Diagnosis of Acute Delirium: A Pilot Study

AHMED AL-HINDAWI[1,2], MARCELA VIZCAYCHIPI[2], AND YIANNIS DEMIRIS[1], (Senior Member, IEEE)

[1]Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.
[2]Department of Anaesthesia, Pain Medicine and Intensive Care, Chelsea and Westminster Hospital NHS Foundation Trust, SW10 9NH London, U.K.

CORRESPONDING AUTHOR: A. AL-HINDAWI (ahmed.al-hindawi@nhs.net)

The work of Ahmed Al-Hindawi and Marcela Vizcaychipi was supported in part by British Medical Association (BMA) Research Fund J Moulton Prize for clinical research into mental health and in part by CW+ Charity. The work of Marcela Vizcaychipi was supported by Westminster Medical School Research Fund.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the London North West NHS Health Research Authority (HRA) and Research and Ethics Committee (REC) under Approval No. 20/LO/0162.

ABSTRACT Objective: Delirium, an acute confusional state, affects 20-80% of patients in Intensive Care Units (ICUs), one in three medically hospitalized patients, and up to 50% of all patients who have had surgery. Its development is associated with short- and long-term morbidity, and increased risk of death. Yet, we lack any rapid, objective, and automated method to diagnose delirium. Here, we detail the prospective deployment of a novel dual-camera contextual eye-tracking platform. We then use the data from this platform to contemporaneously classify delirium. Results: We recruited 42 patients, resulting in 210 (114 with delirium, 96 without) recordings of hospitalized patients in ICU across two centers, as part of a prospective multi-center feasibility pilot study. All recordings made with our platform were usable for analysis. We divided the collected data into training and validation cohorts based on the data originating center. We trained two Temporal Convolutional Network (TCN) models that can classify delirium using a pre-existing manual scoring system (Confusion Assessment Method in ICU (CAM-ICU)) as the training target. The first model uses eye movements only which achieves an Area Under the Receiver Operator Curve (AUROC) of 0.67 and a mean Average Precision (mAP) of 0.68. The second model uses the point of regard, the part of the scene the patient is looking at, and increases the AUROC to 0.76 and the mAP to 0.81. These models are the first to classify delirium using continuous non-invasive eye-tracking but will require further clinical prospective validation prior to use as a decision-support tool. Clinical impact: Eye-tracking is a biological signal that can be used to identify delirium in patients in ICU. The platform, alongside the trained neural networks, can automatically, objectively, and continuously classify delirium aiding in the early detection of the deteriorating patient. Future work is aimed at prospective evaluation and clinical translation.

INDEX TERMS Delirium, eye-tracking, intensive care medicine.

I. INTRODUCTION

DELIRIUM is a clinical syndrome of acutely impaired cognition and memory secondary to a wide spectrum of underlying acute pathologies [1], [2], [3], [4]. It affects between 20-80% of all patients admitted to Intensive Care Unit (ICU), affects up to one in three medically hospitalized patients, and up to one in two of patients who have undergone surgery. It is associated with an increased risk of hospital-acquired infections, increased risk of falls, increased hospital length of stay, and increased cost of stay [5], [6], [7], [8], [9], [10], [11]. The development of delirium also has long-term consequences - following discharge from the hospital, patients who develop delirium have worse cognitive function scores, and this cognitive dysfunction can persist for many years [12], [13], [14].

The diagnosis of delirium in ICU is through the use of the CAM-ICU. This scoring system aims to assess consciousness and attention as surrogate markers of cognition and

IEEE Journal of Translational Engineering in Health and Medicine

memory [15]. However, it requires manual monitoring and awareness from the clinician to trigger the assessment. The scoring system itself is burdensome and can miss episodes of delirium owing to the intermittent nature of manual testing. This can lead to under-diagnosis [16], [17].

Attempts at automating the diagnosis of delirium have been made but have been fraught with difficulties - quantitive Electroencephalogram (EEG) through the use of the BiSpectral Index (BIS) system has been found to correspond to arousal [18], [19]. However, other EEG indices have been used with good metrics [20], [21], [22], [23]. The main drawback of these techniques is that the patient is instrumented and is required to be stationary for the signal to be acquired, which can be difficult as delirious patients can be agitated and combative. Thus, an alternative non-invasive technique that doesn't instrument the patient is needed.

Eye-tracking has been used for the diagnosis and monitoring of neuropsychiatric diseases including schizophrenia, affective disorders, autism spectrum disorder, and Alzheimer's dementia [24], [25], [26], [27]. It has been hypothesized to also be diagnostic for delirium due to the joint role of top-down visual attention modulation and memory encoding of the medial temporal lobe [28], [29]. However, no such platform exists for the eye tracking of patients with delirium, as the eye-tracking device has to meet the requirements of clinical safety owing to the acuity of patients with delirium. Patients with delirium can also be agitated and thus a close wearable eye-tracker is not suitable. Thus, we have developed a bespoke eye-tracking platform that uses a pipeline of neural networks and computer vision algorithms facilitating the acquisition of eye movement and Point of Regard (PoR), the part of the scene the patient is looking at, in real-time, at a safe distance [30], [31], [32]. Despite multiple eye-movement-derived indices being used in neuropsychiatric diseases, such as abnormalities in smooth pursuit in schizophrenia, none have been explored in delirium [26], [33], [34].

To explore eye-tracking for patients suffering from delirium, we developed and validated a novel eye-tracking platform that is suitable for deployment in a clinical environment where it meets criteria for signal acquisition and patient safety [31], [32], [35]. The platform does not require patient level calibration or involvement and does not require a stimulus to be presented and thus is completely non-invasive. In this paper, we describe the prospective multi-center deployment of this camera-based eye-tracking platform for the purpose of ascertaining whether eye-tracking can be used as a biological signal that is diagnostic of delirium.

We use the data gathered from our eye-tracking platform for purposes of classification of delirium using time-series data. We train TCNs on data gathered from one center and validate the models on data gathered from a different center. We train two models - the first uses eye-movement indices only, whilst the second model adds scene contextual information which increases the performance. The result is a model that can take eye-tracking data as input and output a probability of delirium. The resulting models are continuous in nature, can provide a probability of delirium in real-time, and are non-invasive owing to the platform's nature.

II. METHOD

We conducted a prospective multi-center pilot study which was registered on ClinicalTrials.gov[1] for the purpose of the study of eye movements in patients with delirium. It was approved by Health Research Authority (HRA) and Research & Ethics Committee (REC) (Approval number: 20/LO/0162) and was conducted in accordance with the Helsinki declaration. Patients were recruited across Chelsea and Westminster Hospital (CWH) and West Middelsex Hospital (WMH) between November 2020 and February 2022; two interdisciplinary general medical and surgical hospitals. Participants were not remunerated for their participation in the study.

A. PROTOCOL

Following recruitment, patients underwent a once-daily assessment of delirium through CAM-ICU by a singular medically trained intensivist who has had formal training in the diagnosis of delirium. Measurements of eye movements and the fixations on the scene, known as the Point of Regard (PoR), occurred concurrently, for 10 minutes daily until discharge from ICU - see Fig. 1 for an overview of the eye-tracking platform data pipeline [15], [30], [31], [32]. To avoid recruiting many patients who are not delirious, we enriched our cohort by pre-selection avoiding patients at low risk of delirium using the Early PREdiction of DELIRium in ICu patients (E-PRE-DELIRIC) model. To maximize the external validity of the resulting classification models that were trained on the eye-tracking indices, we used data from CWH for training and development while data from WMH was used for external validation. Table 1 demonstrates the patient characteristics across the two sites while Fig. 2 demonstrates the Consolidated Standards of Reporting Trials (CONSORT) diagram for the flow of patients throughout the study.

Consent to enter this study was obtained in one of two ways: as the trial studied delirium, a capacity-losing state, an assessment of capacity was performed, and if present, fully informed consent was gained from the patient directly. Should the participant's mental state, competence, and capacity, mean that they were unable to provide consent, the patient's relatives/friends were approached for advice using the same procedure. In situations where a relative/friend was not available, a nominated consultee was sought which was one of the Medical Consultant Intensivists who were not involved in the care of the patient at the time of advice. Should the patient recover their capacity, fully informed consent was undertaken to ensure that their participation in the study is in line with their wishes.

[1] https://clinicaltrials.gov/ct2/show/NCT04589169

TABLE 1. Characteristics of patients enrolled as part of the deployment of the eye-tracking platform forming a clinical feasibility study. Pertinent confounding variables related to outcomes of interests as well as variables that predict the development of delirium are presented across the two hospitals; Chelsea and Westminster Hospital (CWH) and West Middelsex Hospital (WMH).

	CWH (Development)	WMH (Validation)	Overall
Sessions	186	76	262
CAM-ICU			
Positive	75	39	114
Negative	69	27	96
Unable to measure	42	10	52
Gender			
Female	11	4	15
Male	21	6	27
Age (Years)			
Mean ± Standard Deviation	57.3 ± 19.3	58.0 ± 14.7	57.5 ± 18.1
Frailty Score			
Mean ± Standard Deviation	3.48 ± 1.05	1.50 ± 0.53	3.03 ± 1.27
APACHE-II Probability (%)			
Mean ± Standard Deviation	23.5 ± 14.1	29.5 ± 20.7	24.9 ± 15.7
Urea (mmol/L)			
Mean ± Standard Deviation	10.9 ± 7.08	10.8 ± 5.96	10.9 ± 6.76
Corticosteroids			
No	31	10	41
Yes	1	0	1
Admission Type			
Elective	8	2	10
Emergency	24	8	32
ICU Length of Stay (Days)			
Mean ± Standard Deviation	22.2 ± 25.9	27.3 ± 24.1	23.4 ± 25.2
Unit Outcome			
Alive	25	8	33
Dead	7	2	9

The study's inclusion and exclusion criteria were:

Inclusion Criteria:

1) Aged $\geq$ 18 years
2) Expected risk of delirium as defined by the E-PRE-DELIRIC score to be $\geq$ 20%
3) Expected Length of Stay $\geq$ 2 days

Exclusion Criteria:

1) Lack of Consent
2) Pre-existing diagnosis of Dementia
3) Significant visual impairment
4) Non-concordant eyes
5) The inability for facial recognition and eye tracking to be performed reliably

In relation to exclusion criteria - patients with dementia were excluded as the diagnosis of Delirium-on-Dementia is a clinical challenge and the presence of dementia can confound the diagnosis of delirium; CAM-ICU is specifically not validated for this cohort [36]. The last exclusion criterion 'Patients who were unable to be reliably tracked' was added to ensure that the data processing pipeline can reliably extract the patient's face, detect their landmarks, and regress their gaze vector; an example of a patient that would be excluded would be one undergoing maxillofacial surgery, or with a pre-existing facial deformity. This was added to ensure that the data was of high quality and would generalize to the wider population.

B. EYE TRACKING PLATFORM

Due to the presence of cognitive and memory dysfunction in delirium, traditional eye-tracking platforms cannot be used as they require calibration or are placed within an unsafe distance of the patient. With this motivation, we have previously developed a camera-based non-invasive platform for the continuous measurement of eye signals; specifically head pose, eye horizontal and vertical angles, and blinking status [30], [31], [32]. The patient's environment was instrumented with two cameras connected to a commercially available laptop. One camera was placed at the foot end of the bed facing the patient (termed the head camera) and another was placed behind the patient facing the same direction as the patient's head (termed the scene camera). Compared to other techniques of eye-gaze regression, our method is non-invasive, accurate and precise from a clinically safe distance, and does not limit the patient or restrict clinical care [37]. We deployed this system prospectively in the two centers. The data pipeline of each camera is depicted in Fig. 1.

In summary, the platform performs facial detection and landmark measurement sequentially on images acquired from the head camera using deep convolutional neural networks. These networks were specifically tuned for use in healthcare to minimize false identification of faces and cope with occlusions from medical equipment. The patient's 3D head position and rotation relative to the head camera are then estimated by minimizing the re-projection error of the translation and rotation of a generic pre-specified 3D model onto the measured landmarks. The landmarks also facilitated the extraction of the patient's eye patches which were then used for blink classification and gaze regression using two further neural network ensembles. The platform runs in real-time at a rate of 20 - 30Hz and has been demonstrated to surpass

IEEE Journal of Translational Engineering in Health and Medicine

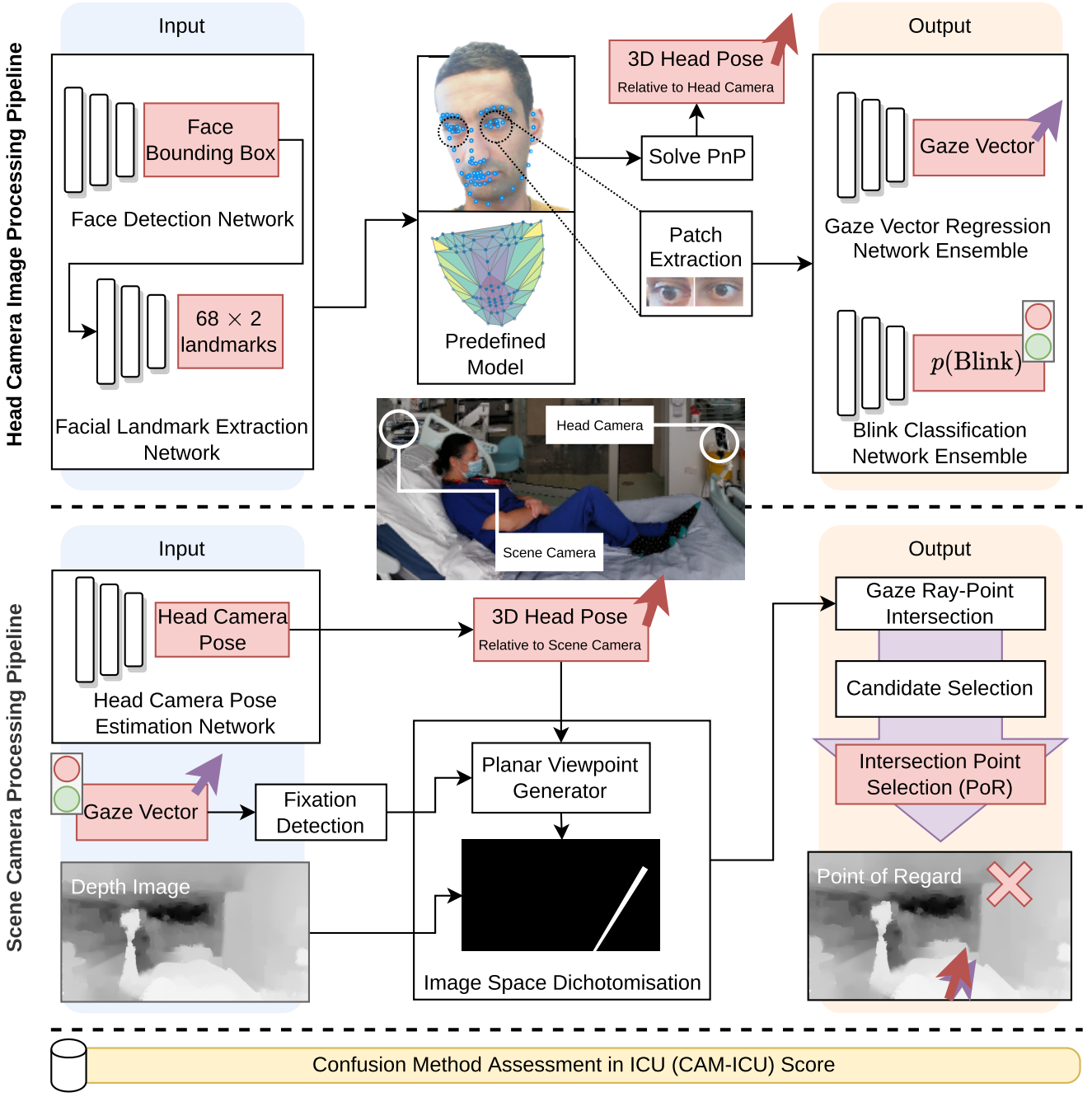


FIGURE 1. Eye tracking pipeline overview. The inset shows the location of the cameras around the bed-space. Red boxes are outputs that are recorded from the pipeline during a session. The head camera locates the patient using a face-detector followed by the extraction of 68 prespecified landmarks which facilitate the calculation of the head-pose of the patient in 3D space relative to the head camera. The landmarks also extract the two eye patches used for blink classification and gaze vector regression. Meanwhile, the scene camera locates the head camera and uses the location of the head camera to transform the 3D head pose into scene camera coordinates. The scene camera, in conjunction with the blink-gated gaze vector, is then used to find the Point of Regard (PoR) - the part of the scene the patient is looking at. The entire pipeline runs at 30Hz. Each recording lasts 10 minutes and is labeled with the current standard, the Confusion Assessment Method in ICU (CAM-ICU) [15], [30].

the required accuracy and precision required for this setting [30], [31], [32]. The gaze vectors were then gathered into fixations which intersected with the scene depth image using a novel image-space gaze-scene intersection algorithm which was developed specifically for this setting and is state-of-the-art [32]. The scene camera also performs pose estimation of the head camera by locating a specialized marker (ChArUco board) located above the camera thus ensuring the patient's measurements are relative to the scene camera. The final outputs of the platform are the patient's gaze vector for each fixation and the PoR in pixel coordinates. Auxiliary measurements are also stored to ensure data validity.

Each 10-minute recording is labeled with CAM-ICU as the ground-truth label for whether the patient was delirious during this recording. To reduce inter-rater variability,

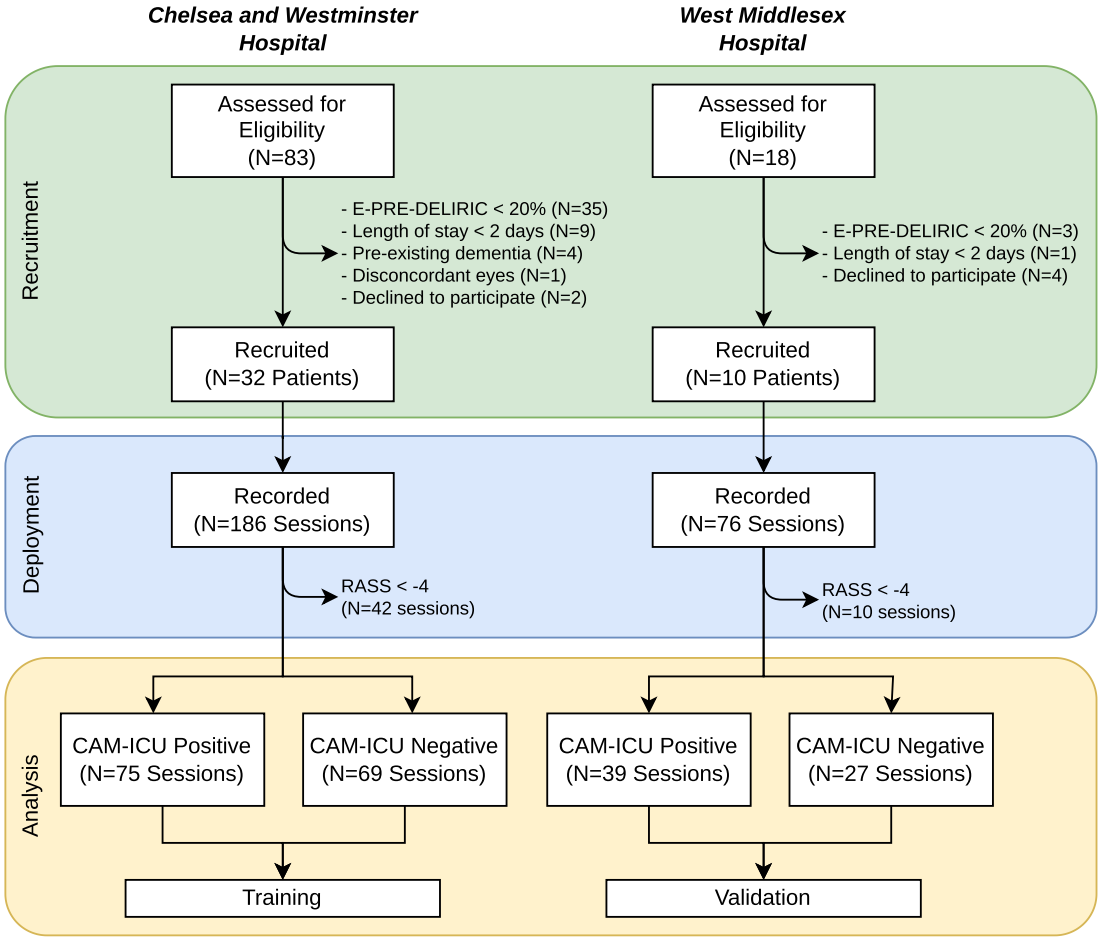


FIGURE 2. Diagram of the flow of patients in the study and how the data was analyzed. Data gathered from Chelsea and Westminster Hospital (CWH) was used for the training while data gathered from West Middelsex Hospital (WMH) was used for validation.

CAM-ICU was conducted by the study personnel rather than the bedside nurse.

C. DATA PROCESSING

Data acquired by the platform are stored in non-ordered sequential storage which is not suitable for machine learning. Thus, an offline processing stage takes the raw data messages and converts them to a format suitable for digestion. The unordered sequential data is thus temporally ordered.

Following the conversion of the measured data into ordered temporal data, a post-processing stage performs data cleaning which removes gaze and scene data where the blink classifier deems that the patient was blinking as per [30]; Fig. 3a illustrates how eye closure/blinking was excluded. Following this, gaze data is converted to fixations based on a dispersion filter. This filter calculates the pair-wise cosine distances across a set of gaze vectors and the inverse cosine of the maximum pair-wise distance is taken to the angle of that set of dispersion, formally:

$$g_v = \{g_1, g_2, \cdots, g_t\} \tag{1}$$

$$dist = 1 - \frac{g_t \cdot g_{t+1}}{\|g_t\|_2 \|g_{t+1}\|_2} \quad \forall (g_t, g_{t+1}) \in g_v \tag{2}$$

$$dispersion = \cos^{-1}(1 - \max(dist)) \tag{3}$$

where $\| * \|_2$ is the ℓ_2 norm of its argument $*$, and $g_t \cdot g_{t+1}$ is the dot product of g_t and g_{t+1}. The result is a dispersion, measured in radians, of the set g_v. The set is composed of gaze vectors where the number of elements in the set comprises a duration of 600ms [38], [39]. If the dispersion of the set g_v is below a threshold, that set is deemed to be a fixation. The dispersion threshold is set to the precision of the gaze tracker at 6° [30].

Following fixation classification, the PoR is measured by the intersection of the fixation vector with the scene as per [32]. This results in a complete dataset per recording session per participant.

D. CLASSIFICATION MODELS

We define the task of delirium classification as a supervised time-series binary classification task. As little is known on what visual and eye-movements features are robust for classification in this task, and to maximize external validity, we opt to use TCNs as they are mechanistically simple but provide powerful classification performance of time-series data [40]. We split our data spatially, data from CWH was used for

IEEE Journal of Translational Engineering in Health and Medicine

training whilst data from WMH was used for validation. We report performance metrics from the validation cohort only.

Two models were trained, the first, aimed at classifying delirium from eye movements, where the inputs to the TCN's X_t are the horizontal and vertical eye angles of each fixation. This model aims to understand if any spatio-temporal eye movements differ from delirious and non-delirious episodes. The second model aims to understand the impact of the scene on eye movements. The part of the scene that the patient is looking at, known as the PoR, was extracted by intersecting the fixation vector into the depth image acquired by the depth camera [32]. Each 224 × 224 patch was then encoded into a 1024-length vector using a ResNet-50 neural network pre-trained on ImageNet and used as the inputs into the TCN [41]. Fig. 3b illustrates the architecture for both models.

III. CLASSIFICATION MODEL SPECIFICATION

As our data is large time series in nature, with variable length, a causal classification technique that can attend to distant time points is required for optimal classification. Whilst both Long-Short Term Memory (LSTM) and TCN can handle arbitrarily long sequences, multi-layer TCN have been shown to be superior to LSTM for long sequences and thus form the basis of our classification models [40].

The TCN acts as a time-series encoder outputting a fixed-length vector that represents the time series in multi-dimensional space. The TCN 's last hidden output thus forms an encoding of the entire time series. This encoding is then fed into a set of fully connected linear layers that output a logit. This logit is then converted to a probability using the so-called softmax function, indicating whether the networks classify the patient's time series as being delirious or not.

As each recording session is 10 minutes in length, rather than using the entire recording session for training, we instead opt to use a window of fixed size within that recording to train the network; this window size was subjected to hyper-parameter search and the result of a window-size of 1000 was used. This has the added benefit of data augmentation as the start of the window can be shuffled thus augmenting the data available for machine learning.

a: DELIRIUM CLASSIFICATION FROM EYE MOVEMENTS MODEL

The TCN for this task is composed of 8 temporal layers followed by 6 fully connected linear layers. Each temporal block consists of a 1-dimensional convolutional filter, with a channel size of 256 and a kernel size of 7 followed by a non-linear activation function (ReLU) and dropout (p=0.05) for regularization.

b: DELIRIUM CLASSIFICATION FROM PoR MODEL

PoR, the part of the scene that the patient is looking at, was extracted by intersecting the fixation vector with the scene image [32]. A crop of the scene, a 224 × 224 patch, was then encoded into a 1024-length vector using a ResNet-52 neural network pre-trained on ImageNet [41]. Thus, the inputs into the TCN (X_t in Fig. 1) are thus the fixed-length vectors.

Similar to the eye movement model specification, the PoR classification model is composed of an 8-layer TCN followed by 5 fully connected linear layers. Each temporal block consists of a 1-dimensional convolutional filter, with a channel size of 512 and a kernel size of 7 followed by a non-linear activation function (ReLU) and dropout (p=0.05) for regularization [40].

TABLE 2. Sensitivities and specificities of the TCN models at different thresholds.

	Eye Movements		Scene	
Threshold	Specificity	Sensitivity	Specificity	Sensitivity
0.1	0.11	0.79	0.18	0.93
0.3	0.48	0.75	0.58	0.82
0.5	0.62	0.71	0.64	0.76
0.7	0.81	0.38	0.85	0.4
0.9	0.89	0.14	0.91	0.14

A. TRAINING REGIME

AdamW optimizer was used with hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a learning rate of 1×10^{-3} [42]. Given that our data is balanced by trial design, non-weighted binary cross-entropy was used for the loss function.

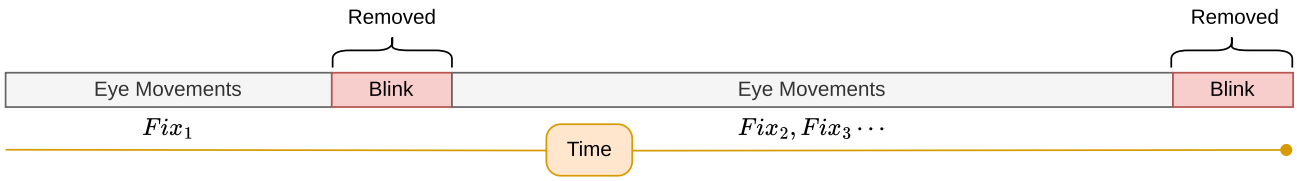
A random search optimization strategy was utilized to find the optimal set of hyper-parameters to maximize performance. A nested cross-validation scheme was used where the training dataset was split into 10 folds where 9 folds were used to train a model with specific hyper-parameters and the last fold was used for validation repeated 10 times and the results averaged [43]. The hyper-parameter optimization random search was bounded within the accepted range of each hyper-parameter. The tuned hyper-parameters were: window-size, TCN layers, channel size, kernel size, number of fully connected layers, learning rate, ℓ_2 regularization, and dropout rate [44], [45]. The best-performing model's parameters were stored and used for validation with the best-validated model used for testing.

B. PERFORMANCE METRICS

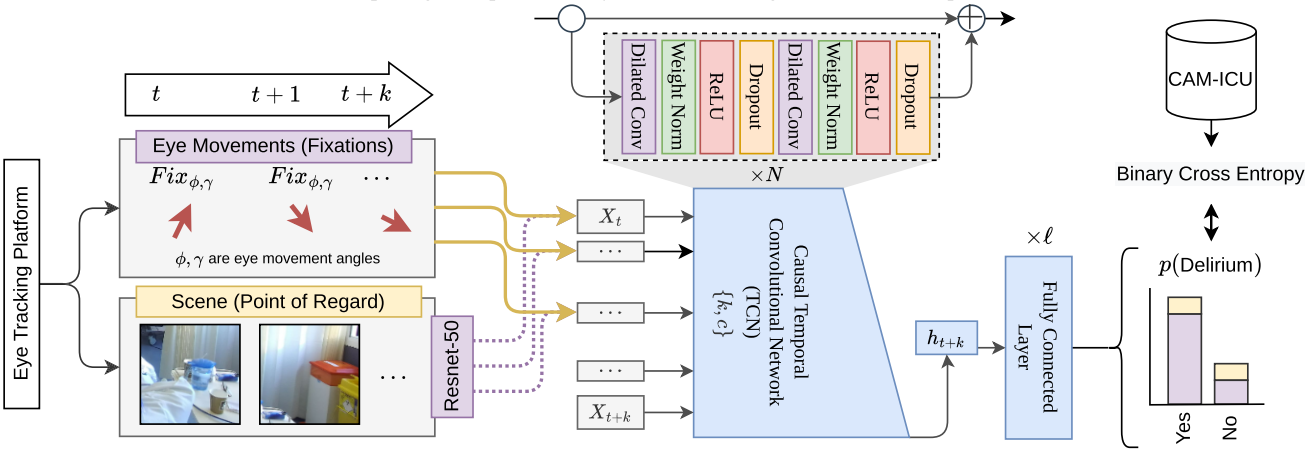
Model performance is demonstrated by using Receiver Operator Characteristic (ROC) on the testing cohort and Precision-Recall Curve (PRC). AUROC demonstrates the accuracy of the model where a number closest to 1.0 indicates perfect classification and 0.5 indicates performance similar to chance. The PRC curve demonstrates the trade-off between Recall and Precision with the mAP summarizing the curve. The closer the mAP is to 1.0 the better the model where 1.0 indicates perfect accuracy compared to a baseline model of 0.

C. CALIBRATION AND THRESHOLD TUNING

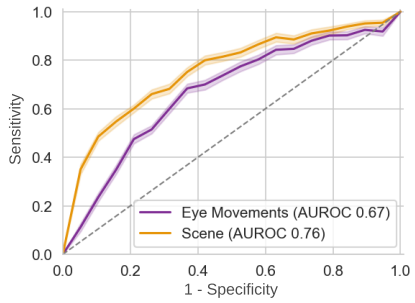
Model calibration was analyzed using normalized calibration curves; these are isotonic curves that compare the estimated binned probabilities to the fraction of observed risk where the diagonal line represents perfect concordance between



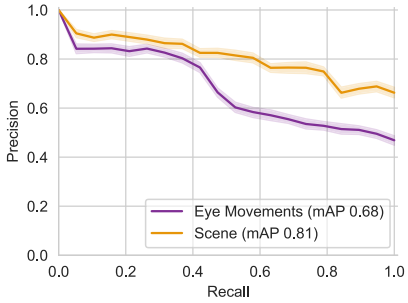
(a) Illustration depicting how periods of eye-closure/blinking were handled as per Section II-C.



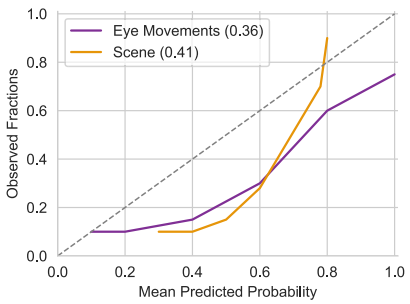
(b) Illustration of the data input into the neural network architecture.



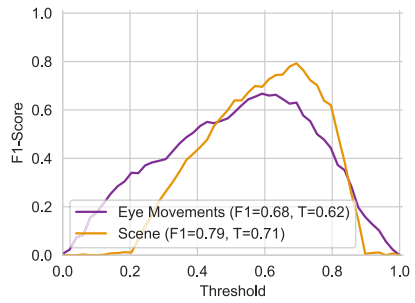
(c) Receiver Operating Characteristic Curve



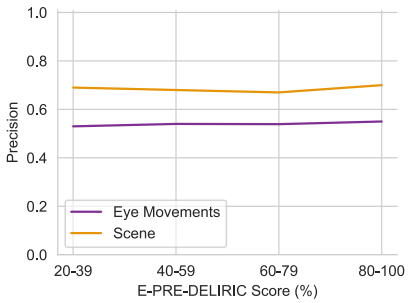
(d) Precision-Recall Curve



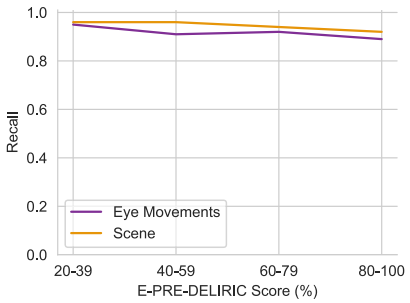
(e) Normalized Calibration Curve



(f) F1-Threshold Tuning Curve



(g) Precision across a range of delirium risk



(h) Recall across a range of delirium risk

FIGURE 3. Illustration of the platform. Following the exclusion of periods of eye closure/blinking and data processing (Fig. 3a), data is then fed into a Temporal Convolutional Network (TCN) which outputs the probability of delirium (Fig. 3b). Two models, one using the time series of eye movements only and the other using a crop of the scene (Point of Regard (PoR)). Fig. 3c and Fig. 3d demonstrate the performance of the trained models for the task of delirium. The eye movements model has good discriminatory performance for the diagnosis of delirium but this performance is improved once scene information is added. Fig. 3e demonstrates calibration typical of neural networks trained using binary cross entropy loss; numbers in brackets represent the Brier calibration score. Fig. 3f demonstrates the F1-score across the sweep of the outcome probability it is being dichotomized.

predictions and observations. Objectively, the Brier score can be used to assess calibration; this is a unit-less score between 0 and 1 where the best-calibrated model attains a score of 0 which is calculated as the sum of residual errors between the prediction and the label [46].

To identify the threshold at which to dichotomize the probabilistic output of the models, we plotted a sweep of the threshold between 0 $\rightarrow$ 1 and calculated the F-1 score, the harmonic mean of the precision and recall at each threshold. The optimal point to dichotomize the probabilistic

IEEE Journal of Translational Engineering in Health and Medicine

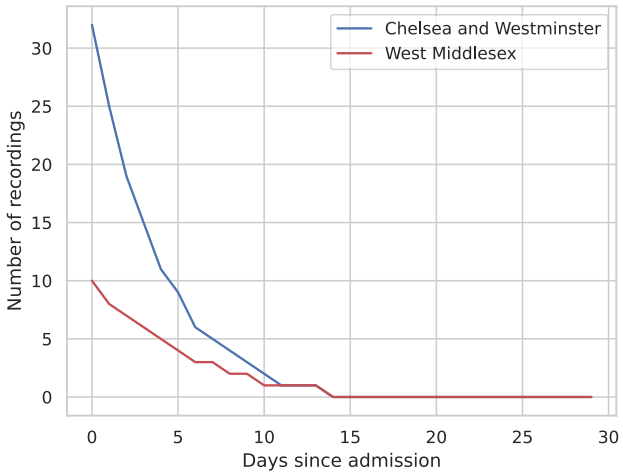


FIGURE 4. Distribution of recordings across the two centers aligned by days since admission. The majority of recordings were from Chelsea and Westminster Hospital (CWH) with a smaller number from West Middlesex Hospital (WMH).

output of the classification models is the threshold which maximizes the F1-score.

IV. RESULTS

The eye-tracking platform facilitated the recruitment of 42 critically unwell patients across two centers, resulting in 262 recording sessions. Table 1 demonstrates the patient characteristics across the two sites, Fig. 2 demonstrates the CONSORT diagram for the flow of patients in the analysis. Recruitment occurred sequentially and all incomers were screened for inclusion and exclusion criteria; the majority of excluded patients were post-operative in nature and thus did not meet the inclusion criteria for length of stay, while a small number of the remaining patients failed to meet the prespecified E-PRE-DELIRIC cutoff of $\geq 20\%$. Only one patient, who suffered from divergent strabismus was excluded from the study due to the platform's inability to infer their direction of gaze. Fig. 4 demonstrates the distribution of recordings across the two centers aligned by days since admission.

Comparing the development and validation cohorts, delirium incidence, as defined by CAM-ICU, was similar ($p = 0.13$, chi-squared test). Similarly, many of the collected confounders, that can lead to the development of delirium, were also found to be similar between the training and validation cohorts ($p = 0.25$, chi-squared test) except the patient's admission urgency ($p < 0.01$, chi-squared test). Following recording, 52 sessions were excluded due to the patient's conscious level corresponding to a Richmond Agitation Sedation Scale (RASS) score ≤ -4 precluding the measurement of delirium using CAM-ICU. No patients were excluded because of the platform's inability to perform eye-tracking. In total, 136,615 fixations were collected that were suitable for analysis.

Each recording session was labeled to originate from a delirious or non-delirious episode by performing manual labeling using CAM-ICU at the time of the recording by the research team to maximize reliability. The data was then split into a training set and a validation set. To maximize external validity, we chose to use data from CWH for model training, and data from WMH for validation. We conducted three studies: the first two concerned the creation of classification models digesting eye tracking data into a probability of delirium; these were trained under a supervised scheme using CAM-ICU as the label. We chose to use neural networks for their powerful ability as universal function approximators without requiring complex specifications, as little is known about the specifics of eye movements in patients with delirium.

A. EYE-MOVEMENTS in DELIRIUM

We first aimed to test whether eye movements themselves, regardless of visual scene attention, were different between episodes of delirium and non-delirium. We trained a model that uses time-series data to classify delirium purely based on the gaze angle. The intuition is that, if the model acts as a universal approximation function, then it should be able to identify the spatio-temporal differences of eye movements if such a difference exists. The trained model (Fig. 3c and 3d) demonstrates adequate discriminatory validation performance with an AUROC of 0.67 and a mAP of 0.68 indicating that eye movements differ between delirious and non-delirious episodes. Table 2 demonstrates the sensitivity and specificity of the model at different thresholds.

B. VISUAL ATTENTION in DELIRIUM

The spatiotemporal eye movement differences outlined in the eye-movements-only model could either be a result of the intrinsic cerebral activity or because of altered processing of the visual information received by the patient. To explore this further, we trained another model that utilizes the contextual information around the patient, namely the scene, for the classification of delirium. The intuition here is that if eye movements are not reactive to the environment, then a model would fail at classifying whether the inputs originated from a delirious episode or a non-delirious episode.

Thus, a crop of the scene that forms the PoR of the participant's fixation was extracted and another TCN model was then trained on the time series of those image crops. This model demonstrates an increase in the discriminatory performance of the classifier to an AUROC of 0.76 and a mAP of 0.81, an 11% increase in the AUROC and a 16% increase in the mAP when compared to using eye movements alone (Fig. 3c and 3d). This suggests that the different eye movements of delirious and non-delirious recordings can be accounted for, at least in part, by extrinsic scene information.

Both trained models exhibit appropriate training diagnostics given the binary cross entropy loss function (Fig. 3e). The classifiers are also stable across a wide range of delirium risk (Figs. 3g and 3h). Table 2 demonstrates the sensitivity and specificity of the model at different thresholds.

V. DISCUSSION

Delirium is common and affects a wide range of patients with profound short and long-term consequences. Yet, no objective

marker of been developed. While visual attention has been hypothesized to be diagnostic for delirium, testing this hypothesis has been hampered by a lack of a technological solution that is clinically safe, accurate, precise, and meets empiric requirements. In our previous work, we developed and validated a state-of-the-art eye-tracking platform suitable for the continual non-invasive eye-tracking of delirious patients across two hospitals [30], [31], [32], [35]. In this manuscript, we sought to understand the utility of eye-tracking for the classification of delirium using our platform.

Using the data acquired using the platform from two general medical and surgical ICUs, we conducted two studies. The first study was aimed at understanding eye movement characteristics between delirium episodes and non-delirium episodes. A TCN, took blink-gated fixation adjusted gaze vectors as inputs and predicted delirium. The classification accuracy on a validation dataset suggested that there are differences in the fixation angle behavior between delirious and on-delirious episodes with good classification accuracy indicating that fixation-angle is separable between patients suffering with and without delirium. To address whether this difference is due to internal or external factors, we trained another TCN – if eye movements were intrinsic - i. e.originating from internal mechanisms without any external influence, then a classifier trained on scene information would not be able to discriminate between delirious and non-delirious episodes. This scene TCN found that scene information, through the extraction of PoRs, increased classification accuracy suggesting that scene information contributes to the eye movements characteristics of delirium. By adding PoR as contextual information, the classifier can more accurately delineate between patients suffering from delirium compared to those who are not. This finding the first of its kind to shed light on whether visual attention can discriminate episodes of delirium and non-delirium. The performance of the classifiers was stable across a wide range of delirium risk and the classifiers were well-calibrated but the best-performing model only achieved an AUROC of 0.76 and a mAP of 0.81. This is a good performance but not yet at the level of clinical utility.

Our study had several strengths. Firstly, it is the first study of its kind that looks at eye-tracking in critically unwell patients. We prospectively gathered eye-tracking data from patients across two centers where the first center's data was used to create and train the classifier while the second center's data was used for validation. This provides assurances on the validity of our findings of delirium classification. Secondly, the CAM-ICU test was conducted by the same personnel across both sites decreasing the inter-rater variability and increasing diagnostic confidence. Thirdly, the nature of the dual-camera solution provides a clinically safe system that is non-invasive and hands-free enabling continuous care without any instrumentation, unlike eye-tracking glasses. This makes it deployable across many healthcare institutions including medical and surgical wards, as well as community nursing homes – ICU was chosen for its large concentration of delirious patients as part of a pilot feasibility trial. We envisage that part of the deployment of the system would result in the fixation of the cameras in the patient's bedspace thus making them a fixture that does not require regular set-up.

Putting the results of this study in a clinical context, we found that eye movements, and specifically, where in the scene the patient is paying attention to, can readily diagnose delirium in an automated way. This provides an objective marker that is free from the moderately high inter-observer variability of CAM-ICU. This finding is in keeping with the current thinking around delirium where visual attention is thought to be a key diagnostic feature relating to the interplay between working memory and visual processing [47]. This is also a useful biological signal that is the direct result of a cognitive process in an acutely ill patient and can serve as a foundation to build further clinical translational work including the development of a clinical decision tool, the understanding of the neurological basis of visual inattention in delirium, and the development of a clinical trial to understand the impact of automatic delirium monitoring on patient outcomes.

A. LIMITATIONS

While the study we conducted has several strengths, we wish to highlight some limitations of our approach, ways in which we envisage they could be addressed, and future work in this area.

Firstly, the camera system is required to be positioned at a place where the eyes of the patient can be viewed. This limits the usability of the system to supine patients – *i. e.* not lying on their front, and not significantly lying on their side. It is a standard of care in ICU to nurse patients at 30° head-up in a supine position to minimize aspiration risk and thus the supine forward-facing position encompasses the majority of patients. Active Vision, the research area involved in finding the optimal position of the cameras to maximize signal acquisition could provide a potential solution for patients lying on their side. Similarly, occlusion of the patient's eyes by clinical staff, or other objects, is another limitation. Fig. 1 illustrates that the first stage is facial detection, which if occluded, would fail thus stopping the eye tracking pipeline from progressing. Taking this limitation to an extreme where the majority of the time, the view is obstructed, would result in our system being an intermittent test, similar to the current diagnostic standard of CAM-ICU.

B. FUTURE WORK

This pilot feasibility study looking at eye-tracking in ICU could be advanced further in several directions, both from an engineering perspective and a clinical translational perspective. We firstly wish to investigate the classifiers in a prospective setting where their clinical utility would be scrutinized, and secondly, to focus on understanding the neurological basis of visual inattention in delirium.

1) PROSPECTIVE EVALUATION

To achieve our first goal, the classification models, which take the eye-tracking data as input and result in a diagnostic probability of delirium, would require further work to increase their performance and prospective evaluation prior to its use as part of a clinical decision tool. Formal decision analysis will also required to ascertain the impact of automatic delirium monitoring across various thresholds as well as qualitative work relating to appropriate threshold tuning prior to informing of a positive delirium diagnosis. Other diagnostic metrics can also be evaluated, such as time-to-diagnosis, important patient outcomes following intervention against the current standard, as well as potential harm from false positive diagnoses. This will also facilitate the understanding of the technology iteratively, improving the technology alongside clinical outcomes. In addition to this, appropriate regulatory approvals (e. g.from the Food & Drug Administration in the USA and the Medicine Healthcare products Regulatory Agency in the UK) will be sought to ensure the technology is safe and effective for use in the clinical setting.

2) NEUROLOGICAL BASIS of VISUAL IN-ATTENTION

For our second goal, one of understanding the neurological basis of visual inattention, this can either be through EEG recordings of episodes of delirium aiming at the inspection of the interplay between the medial temporal lobe and the ventral visual processing stream. Alternatively, simulations of the visual attention, given the scene, can then be compared to the actual visual attention in episodes of delirium - thus contrasting healthy and delirious minds under a cognitive architecture scheme. We would also wish to phenotype delirium using the eye-tracking platform and then compare the phenotypes to the underlying neurological basis of delirium. This would provide a deeper understanding of the neurological basis of delirium and provide a foundation for the development of novel treatments. Unfortunately, the current study was not designed to answer these questions, with a limited set of patients with delirium, and thus future work is required to address these questions.

3) PHENOTYPING DELIRIUM USING EYE-TRACKING

An interesting direction of future work would be to phenotype the type of delirium based on our eye-tracking classifiers. Beyond hypo-active, and hyper-active delirium which are clinically obvious once delirium is diagnosed, the underlying cause is often not immediately known and clinicians often correlate the cause with the patient's current state as well as their underlying diagnosis - *e. g.* inflammatory delirium if the patient has a concurrent infection, or metabolic delirium if the patient has electrolyte disturbances.

However, the etiology of delirium can often be different from the underlying disease - e. g.metabolic delirium owing to sepsis-induced hepatitis while the patient has an infection. Thus, hinting at the cause of delirium would lead to different treatments and is of importance for future work.

Many techniques exist in the literature that can reveal the underlying groups without further data - e. g.self-supervised techniques would use contrastive techniques to 'pull' data points that are similar to each other, and 'push' data points that are disparate from each other during the training phase [48]. These techniques require a significant number of patients as the phenotyping would have to be at the level of the patient and not the recording. Simpler techniques can be Principal Component Analysis (PCA), or T-Distributed Stochastic Neighbor Embedding (TSNE) [49], [50].

An alternative technique, linear probing, would place a linear layer on one of the hidden layers of the neural network, and only that layer would then be trained using supervised learning for a classification task. This would require the *a-priori* knowledge of the number of clusters that can output a probability of belonging to a pre-specified number of classes [51]. However, again, this technique would require a larger sample size to work effectively.

VI. CONCLUSION

Delirium affects a wide range of patients with severe consequences. Yet, an objective automated system has not been developed. We demonstrated how eye tracking, as performed in a non-invasive, calibration-free manner, can automatically classify delirium in Intensive Care Unit (ICU) to good performance metrics in a pilot feasibility study. Future work is aimed at improving the performance of the models, validating the classification models for clinical use, and using the biological signal to phenotype delirium to understand the neurological basis of this signal.

ACKNOWLEDGMENT

The authors would like to thank the patients, families, and members of staff at Chelsea & Westminster Hospital (CWH) and West Middlesex Hospital (WMH) for their contributions to this work. Ahmed Al-Hindawi and Marcela Vizcaychipi would like to thank the British Medical Association (BMA) Research Fund J Moulton Prize for clinical research into mental health and CW+ charity for their funding of this project. Ahmed Al-Hindawi would like to thank members of the Personal Robotics Laboratory, Singer Lab, and UCL HAL Lab. Specifically, Professor Mervyn Singer, Dr. Nishkantha Aruklumaran, Dr. Steve Harris, Dr. Timothy AC Snow, Dr. Tobias Zimmermann, Dr. Rodrigo Chacon-Quesada, Dr. Tobias Fischer, and Dr. Joshua Elsdon. Marcela Vizcaychipi would like to thank the Westminster Medical School Research Fund for its on going support. Yiannis Demiris would like to thank the Royal Academy of Engineering for his chair in Emerging Technologies.

COMPETING INTERESTS

None declared by all authors

REFERENCES

[1] J. E. Wilson et al., "Delirium," *Nature Rev. Disease Primers*, vol. 6, no. 1, pp. 1–26, Nov. 2020.

[2] R. Palmu, "Mental disorders among burn patients," *Burns*, vol. 36, no. 7, pp. 1072–1079, 2011.

[3] M. Aldemir, S. Özen, I. H. Kara, A. Sir, and B. Baç, "Predisposing factors for delirium in the surgical intensive care unit," *Crit. Care*, vol. 5, no. 5, p. 265, Sep. 2001.
[4] M. Lundström et al., "Postoperative delirium in old patients with femoral neck fracture: A randomized intervention study," *Aging Clin. Experim. Res.*, vol. 19, no. 3, pp. 178–186, Jun. 2007.
[5] N. C. Andersen-Ranberg, "Haloperidol for the treatment of delirium in ICU patients," *New England J. Med.*, vol. 387, no. 26, pp. 2425–2435, Dec. 2022.
[6] S. Ouimet, B. P. Kavanagh, S. B. Gottfried, and Y. Skrobik, "Incidence, risk factors and consequences of ICU delirium," *Intensive Care Med.*, vol. 33, no. 1, pp. 66–73, Jan. 2007.
[7] D. L. Leslie and S. K. Inouye, "The importance of delirium: Economic and societal costs," *J. Amer. Geriatrics Soc.*, vol. 59, no. S2, pp. S241–S243, Nov. 2011.
[8] E. Ely et al., "The impact of delirium in the intensive care unit on hospital length of stay," *Intensive Care Med.*, vol. 27, no. 12, pp. 1892–1900, Dec. 2001.
[9] S.-M. Lin et al., "The impact of delirium on the survival of mechanically ventilated patients," *Crit. Care Med.*, vol. 32, no. 11, pp. 2254–2259, 2004.
[10] J. McCusker, M. Cole, M. Abrahamowicz, F. Primeau, and E. Belzile, "Delirium predicts 12-month mortality," *Arch. Internal Med.*, vol. 162, no. 4, p. 457, Feb. 2002.
[11] G. A. Caplan, A. Teodorczuk, J. Streatfeild, and M. R. Agar, "The financial and social costs of delirium," *Eur. Geriatric Med.*, vol. 11, no. 1, pp. 105–112, Feb. 2020.
[12] J. C. Jackson, S. M. Gordon, R. P. Hart, R. O. Hopkins, and E. W. Ely, "The association between delirium and cognitive decline: A review of the empirical literature," *Neuropsychol. Rev.*, vol. 14, no. 2, pp. 87–98, Jun. 2004.
[13] Z. J. Kunicki et al., "Six-year cognitive trajectory in older adults following major surgery and delirium," *JAMA Internal Med.*, vol. 183, no. 5, p. 442, May 2023.
[14] C. B. Mortensen et al., "Long-term outcomes with haloperidol versus placebo in acutely admitted adult ICU patients with delirium," *Intensive Care Med.*, vol. 50, no. 1, pp. 103–113, Jan. 2024.
[15] E. W. Ely et al., "Delirium in mechanically ventilated patients: Validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU)," *JAMA*, vol. 286, no. 21, p. 2703, Dec. 2001.
[16] J.-D. Gaudreau, P. Gagnon, F. Harel, A. Tremblay, and M.-A. Roy, "Fast, systematic, and continuous delirium assessment in hospitalized patients: The nursing delirium screening scale," *J. Pain Symptom Manage.*, vol. 29, no. 4, pp. 368–375, Apr. 2005.
[17] M. A. Pisani, K. L. B. Araujo, P. H. Van Ness, Y. Zhang, E. W. Ely, and S. K. Inouye, "A research algorithm to improve detection of delirium in the intensive care unit," *Crit. Care London, England*, vol. 10, no. 4, Aug. 2006.
[18] H. Koponen, J. Partanen, A. Pääkkönen, E. Mattila, and P. J. Riekkinen, "EEG spectral analysis in delirium," *J. Neurol., Neurosurgery, Psychiatry*, vol. 52, no. 8, pp. 980–985, Aug. 1989.
[19] K. Plaschke et al., "Early postoperative delirium after open-heart cardiac surgery is associated with decreased bispectral EEG and increased cortisol and interleukin-6," *Intensive Care Med.*, vol. 36, no. 12, pp. 2081–2089, Aug. 2010.
[20] A. W. van der Kooi, F. S. S. Leijten, R. J. van der Wekken, and A. J. C. Slooter, "What are the opportunities for EEG-based monitoring of delirium in the ICU?" *J. Neuropsychiatry Clin. Neurosciences*, vol. 24, no. 4, pp. 472–477, Jan. 2012.
[21] A. Hunter, B. Crouch, N. Webster, and B. Platt, "Delirium screening in the intensive care unit using emerging QEEG techniques: A pilot study," *AIMS Neurosci.*, vol. 7, no. 1, pp. 1–16, 2020.
[22] H. Sun et al., "Automated tracking of level of consciousness and delirium in critical illness using deep learning," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–8, Sep. 2019.
[23] M. Pollak, S. Leroy, V. Röhr, E. N. Brown, C. Spies, and S. Koch, "EEG biomarkers from anesthesia induction to identify vulnerable patients at risk for postoperative delirium," *Anesthesiology*, vol. 140, no. 5, pp. 979–989, Jan. 2024, doi: 10.1097/ALN.0000000000004929.
[24] S. A. Beedie, D. M. St. Clair, and P. J. Benson, "Atypical scanpaths in schizophrenia: Evidence of a trait- or state-dependent phenomenon?" *J. Psychiatry Neurosci.*, vol. 36, no. 3, pp. 150–164, May 2011.
[25] C. M. Loughland, L. M. Williams, and E. Gordon, "Schizophrenia and affective disorder show different visual scanning behavior for faces: A trait versus state-based distinction?" *Biol. Psychiatry*, vol. 52, no. 4, pp. 338–348, Aug. 2002.
[26] P. Trillenberg, R. Lencer, and W. Heide, "Eye movements and psychiatric disease," *Current Opinion Neurol.*, vol. 17, no. 1, pp. 43–47, Feb. 2004.
[27] Y. Zhang, T. Wilcockson, K. I. Kim, T. Crawford, H. Gellersen, and P. Sawyer, "Monitoring dementia with automatic eye movements analysis," in *Intelligent Decision Technologies*, vol. 57, I. Czarnowski, A. M. Caballero, R. J. Howlett, and L. C. Jain, Eds. Cham, Switzerland: Springer, 2016, pp. 299–309.
[28] L. R. Squire, C. E. Stark, and R. E. Clark, "The medial temporal lobe," *Annu. Rev. Neurosci.*, vol. 27, pp. 279–306, Jul. 2004.
[29] C. Exton and M. Leonard, "Eye tracking technology: A fresh approach in delirium assessment?" *Int. Rev. Psychiatry*, vol. 21, no. 1, pp. 8–14, Jan. 2009.
[30] A. Al-Hindawi, M. P. Vizcaychipi, and Y. Demiris, "Continuous non-invasive eye tracking in intensive care," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1869–1873.
[31] A. Al-Hindawi, M. Vizcaychipi, and Y. Demiris, "Faster, better blink detection through curriculum learning by augmentation," in *Proc. Symp. Eye Track. Res. Appl.* New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 1–7.
[32] A. Al-Hindawi, M. P. Vizcaychipi, and Y. Demiris, "What is the patient looking at? Robust gaze-scene intersection under free-viewing conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2430–2434.
[33] L. A. Abel, S. Levin, and P. S. Holzman, "Abnormalities of smooth pursuit and saccadic control in schizophrenia and affective disorders," *Vis. Res.*, vol. 32, no. 6, pp. 1009–1014, Jun. 1992.
[34] P. S. Holzman, L. R. Proctor, and D. W. Hughes, "Eye-tracking patterns in schizophrenia," *Science*, vol. 181, no. 4095, pp. 179–181, Jul. 1973.
[35] A. Al-Hindawi and M. Vizcaychipi, *Continuous Non-Invasive Eye Tracking for the Early Detection of Delirium on the Intensive Care Unit*, document NCT04589169, Clin. Trial Registration, Dec. 2020.
[36] D. M. Fick, J. V. Agostini, and S. K. Inouye, "Delirium superimposed on dementia: A systematic review," *J. Amer. Geriatrics Soc.*, vol. 50, no. 10, pp. 1723–1732, Oct. 2002.
[37] H. Ahn, J. Jeon, D. Ko, J. Gwak, and M. Jeon, "Contactless real-time eye gaze-mapping system based on simple Siamese networks," *Appl. Sci.*, vol. 13, no. 9, p. 5374, Apr. 2023.
[38] D. E. Irwin, "Memory for position and identity across eye movements," *J. Experim. Psychol., Learn., Memory, Cognition*, vol. 18, no. 2, pp. 307–317, 1992.
[39] X. P. Kotval and J. H. Goldberg, "Eye movements and interface component grouping: An evaluation method," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, Oct. 1998, vol. 42, no. 5, pp. 486–490.
[40] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[43] M. W. Browne, "Cross-validation methods," *J. Math. Psychol.*, vol. 44, no. 1, pp. 108–132, Mar. 2000.
[44] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[45] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, 2016, pp. 1019–1027.
[46] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, Jan. 1950.
[47] D. Collerton, E. Perry, and I. McKeith, "Why people see things that are not there: A novel perception and attention deficit model for recurrent complex visual hallucinations," *Behav. Brain Sci.*, vol. 28, no. 6, pp. 737–757, Dec. 2005.
[48] P. Khosla et al., "Supervised contrastive learning," 2020, *arXiv:2004.11362*.
[49] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987.
[50] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Commun.*, vol. 10, no. 1, p. 5415, Nov. 2019.
[51] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2016, *arXiv:1610.01644*.