# scientific reports

Check for updates

OPEN

# Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network

Chin Wen Tan[1,2✉], Tiehua Du[3], Jing Chun Teo[4], Diana Xin Hui Chan[2,5], Wai Ming Kong[3] & Ban Leong Sng[1,2]

Self-reported pain scores are often used for pain assessments and require effective communication. On the other hand, observer-based assessments are resource-intensive and require training. We developed an automated system to assess pain intensity in adult patients based on changes in facial expression. We recruited adult patients undergoing surgery or interventional pain procedures in two public healthcare institutions in Singapore. The patients' facial expressions were videotaped from a frontal view with varying body poses using a customized mobile application. The collected videos were trimmed into multiple 1 s clips and categorized into three levels of pain: no pain, mild pain, or significant pain. A total of 468 facial key points were extracted from each video frame. A customized spatial temporal attention long short-term memory (STA-LSTM) deep learning network was trained and validated using the extracted keypoints to detect pain levels by analyzing facial expressions in both the spatial and temporal domains. Model performance was evaluated using accuracy, sensitivity, recall, and F1-score. Two hundred patients were recruited, with 2008 videos collected for further clipping into 10,274 1 s clips. Videos from 160 patients (7599 clips) were used for STA-LSTM training, while the remaining 40 patients' videos (2675 clips) were set aside for validation. By differentiating the polychromous levels of pain (no pain versus mild pain versus significant pain requiring clinical intervention), we reported the optimal performance of STA-LSTM model, with accuracy, sensitivity, recall, and F1-score all at 0.8660. Our proposed solution has the potential to facilitate objective pain assessment in clinical settings through the developed STA-LSTM model, enabling healthcare professionals and caregivers to perform pain assessments effectively in both inpatient and outpatient settings.

**Keywords** Facial, Pain assessment, Machine learning, Artificial intelligence, Deep learning

**Abbreviations**
AFER      Automated facial expression recognition
AI      Artificial intelligence
ASA      American society of anesthesiologists
AUC      Area under the curve
BMI      Body mass index
CEN      Committee for standardization
CNN      Convolution-neural-network

[1]Department of Women's Anesthesia, KK Women's and Children's Hospital, 100 Bukit Timah Road, Singapore 229899, Singapore. [2]Anesthesiology and Perioperative Sciences Academic Clinical Program, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore. [3]Biomedical Engineering and Materials Group, Nanyang Polytechnic, 180 Ang Mo Kio Avenue 8, Singapore 569830, Singapore. [4]Digital Integration Medical Innovation and Care Transformation, KK Women's and Children's Hospital, 100 Bukit Timah Road, Singapore 229899, Singapore. [5]Department of Anesthesiology, Singapore General Hospital, SingHealth Tower, 10 Hospital Boulevard #19-01, Singapore 168582, Singapore. ✉email: tan.chin.wen@kkh.com.sg

| CPOT | Critical-care pain observation tool |
|------|-------------------------------------|
| DL | Deep learning |
| EDLM | Ensemble deep learning framework |
| EMR | Electronic medical record |
| FDA | Food and drug administration |
| FN | False negative |
| FP | False positive |
| GPUs | Graphics processing units |
| IMDRF | International medical device regulators forum |
| IQR | Interquartile range |
| ISO | International organization for standardization |
| KKH | KK women's and children's hospital |
| LSTM | Long short-term memory |
| MDR | Medical device regulation |
| MIntPAIN | Multimodal intensity pain |
| MOH | Ministry of health |
| NRS | Numerical rating scale |
| PC | Personal computer |
| RAM | Random-access memory |
| RNN | Recurrent neural network |
| ROC | Receiver-operating curve |
| SaMD | Software as a medical device |
| SD | Standard deviation |
| SGH | Singapore general hospital |
| SMOTE | Synthetic minority over-sampling technique |
| STA | Spatial temporal attention |
| STA-LSTM | Spatial temporal attention long short-term memory |
| SVD | Singular value decomposition |
| TN | True negative |
| TP | True positive |
| TRIPOD | Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis |
| TS | Technical specification |
| UNBC | University of Northern British Columbia |
| VAS | Visual analogue scale |

Pain is an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage. Poorly controlled pain after surgery is associated with both short- and long-term adverse impacts on patient outcomes, including delayed recovery and an increased risk of chronic pain[1].

Conversely, excessive opioid dosing may lead to adverse effects such as sedation and respiratory depression which may increase the risk of patient morbidity. Hence, optimal pain management requires regular, timely, and accurate assessment of pain intensity to ensure that the appropriate amount of analgesia is administered[2].

Currently, the reference standard for pain assessment is patient-reported pain scores, e.g., numerical rating scale (NRS) or visual analogue scale (VAS)[3]. However, self-reporting of pain can be challenging for patients with cognitive dysfunction or those who are non-communicative. Observer-based pain assessment serves as an alternative method, where the observer estimates pain intensity using visual and physical cues. Nevertheless, this method is often associated with several limitations: (1) assessments are resource-intensive and increase workload and cost[4]; (2) the accuracy and reproducibility of pain assessments may be affected by observers' cognitive biases, as well as by the patients' demographic and psychological factors, and previous pain experiences; and (3) some pain scales are complex and involve live measurements/calculations, requiring specific training and precluding their use in non-clinical settings[5]. Common facial expressions associated with pain include the combination of the following facial movements: facial grimacing, brow furrowing, eye squeezing or closing, cheek raising, open mouth or clenched teeth and nasal flaring or wrinkling[4,6,7]. As a result, there could be significant variability in pain assessment by patient observation.

Machine learning and automated facial expression recognition (AFER) have been developed to provide objective and cost-effective pain assessment in several healthcare settings where patient-reported pain scores may not be feasible[4]. These modalities rely on facial expressions that signify pain and adopt a framework comprising four modules: face detection, alignment, feature extraction, and classification[8]. For instance, PainChek, an AFER system, is used in Australian hospitals to assist with pain assessments in patients with dementia[9,10]. However, PainChek has been validated only in Caucasian patients with dementia, which may limit its applicability to the local Asian population and other clinical settings. Due to ethical concerns, the development of AFER systems often begins with videotaping cognitively healthy adults experiencing acute pain or other distressing states, as this provides a better controlled experimental setting[11]. Hence, we aimed to develop an automated system using machine learning to assess the pain intensity through changes in facial expression in Asian adult patients undergoing surgery or interventional pain procedures.

## Methods
### Study setting
This study was conducted at two healthcare institutions in Singapore: KK Women's and Children's Hospital (KKH) and Singapore General Hospital (SGH) between May 2022 and December 2023. The study received

approval from the SingHealth Centralized Institutional Review Board (Reference Number: 2019/2293) on 26 Apr 2019 and was registered on Clinicaltrials.gov (NCT04011189) on 8 Jul 2019. Written informed consent was obtained from all participants and the study is conducted in accordance with the Declaration of Helsinki. This article conforms to the relevant transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines[12].

The study population included adult patients undergoing surgery or interventional pain procedures in both inpatient and outpatient clinic settings, with an American Society of Anesthesiologists (ASA) physical status I-III, and aged between 21 and 70 years old. Pregnant patients, as well as those with medical conditions, including psychiatric disorders (e.g., anxiety, depression), neurological disorders (e.g., cerebrovascular accident, Parkinson's disease) and musculoskeletal resulting in facial abnormalities or restrictions, were excluded from the study.

Demographic data was collected upon recruitment (Fig. 1), and the recruited patients were asked to complete a questionnaire on general health status (EQ-5D-3L) and to rate their NRS pain score (0 being no pain, 10 being worst pain imaginable)[13]. Their facial expressions and body pose from a frontal view were videotaped by a trained study team member using a customized mobile application that utilized a dialog tree to guide the videotaping based on five specific actions: sitting, taking a deep breath (holding for three counts before exhaling), sitting-to-standing, standing, and standing-to-sitting (Fig. 2). All videos were recorded in high definition (2160×3840) in MPEG-4 format with H.264 coding at 30 frames per second (fps). The video duration depended on the individual performing the specific actions. The videotaping session took about 10–20 min, including pain-related questions where patients were asked to rate their pain score after performing the specific actions.

Patients underwent surgery or procedures performed as per routine clinical practice. After surgery, they were reviewed at post-operative day -1 and/or -2 before discharge, and asked to rate their pain scores and have their facial expressions and body pose videotaped from a frontal view. Similarly, those who underwent interventional pain procedures received routine clinical care and were asked to rate their pain scores and undergo videotaping either after the procedure or during their next consultation visit.

## Data processing

The collected videos were trimmed into multiple 1-s clips and categorized based on observer pain ratings, with the criteria modified from critical-care pain observation tool (CPOT) to focus on the changes in facial expression[14,15]. Three levels of pain were defined as follows: i) no pain if the patient was in a relaxed, natural manner with no muscle tension observed in the facial features; ii) mild pain if the patient was tensed, with facial features such as levator contraction, frowning, orbit tightening etc.; and iii) significant pain if the patient was grimacing and exhibited all the features in ii), along with tightly closed eyelids. This definition is considered clinically relevant as these levels typically correspond to NRS 0 (no pain), 1–3 (mild pain), and 4–10 (significant pain). Patients who score 4 and above would typically require additional intervention to relieve pain as per clinical practice including the study sites[16].

We further processed the videos by extracting the facial key points from each frame to maintain patient confidentiality, ensuring no identifiable parameters were linked during the analysis. A total of 468 facial key
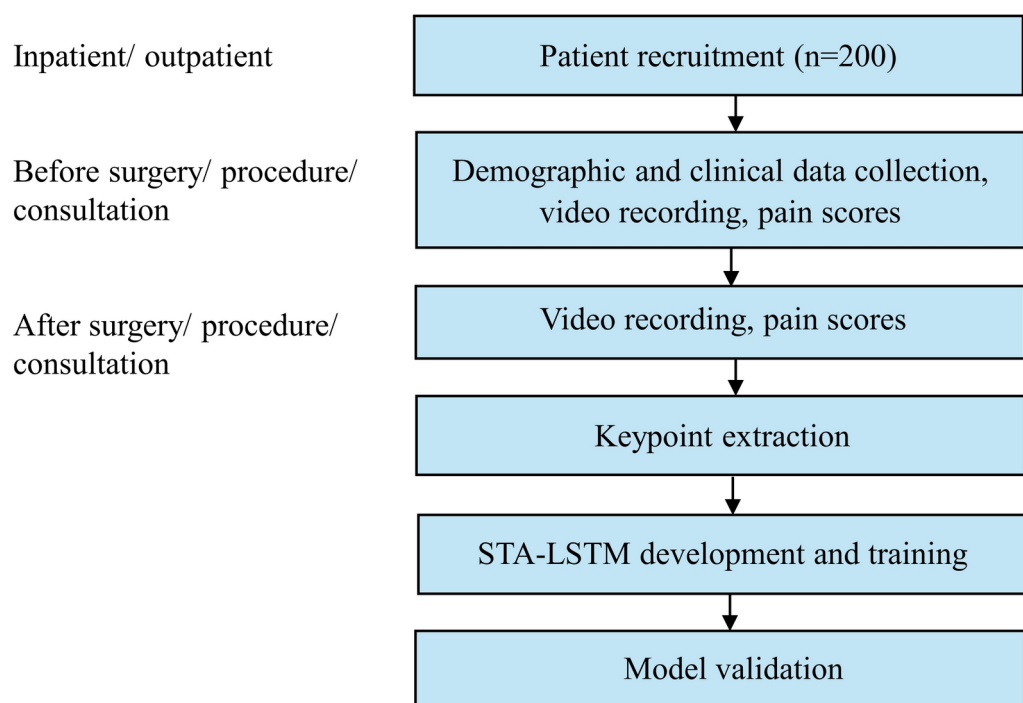


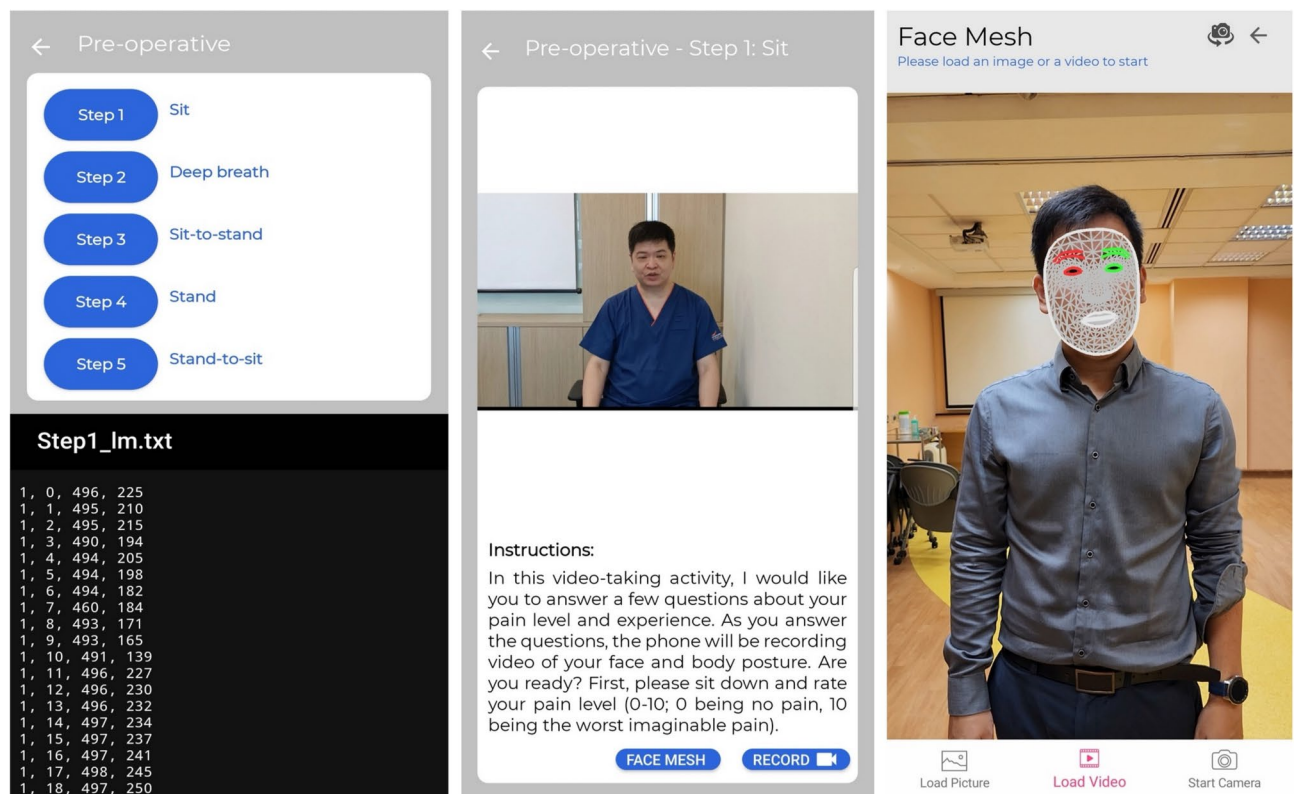**Fig. 1.** Study workflow. STA-LSTM, spatial temporal attention long short-term memory.

**Fig. 2**. User interface of customized mobile application for videotaping, and the keypoints extracted from the video.

points were extracted from each frame using MediaPipe, covering critical facial regions such as eyebrows, eyes, nose, lips, and jawline. Hence, a 1-s video with a frame rate of 30 fps would contain 14,040 key points. In addition, the patients' videos were taken at different location, angle, and scale, which may affect the quality of the extracted facial key points. Therefore, we implemented a 3D normalization algorithm to account for variations in video recordings caused by differences in camera setups, distances, head dimension and positioning. This algorithm ensured consistency in the positioning, scaling, orientation of 468 facial landmarks, and the isolation of facial expression movements for analysis[17,18]. The process began with similarity transformations comprising rotation, translation, and scaling using facial landmarks that are more stable and resistant to deformation: left lateral canthus (point 226), right lateral canthus (point 446), and subnasale (point 2), all of which were mapped to fixed target points to ensure uniform scale and a front-facing orientation[19]. The similarity transformation was then determined by aligning the source and target points around their centroids, computing a scaling factor based on the average distances between corresponding points, and applying this factor to the source data. Next, a covariance matrix was generated from the centered points, with Singular Value Decomposition (SVD) utilized to extract the rotation matrix while accounting for possible reflections[20]. The final translation vector was derived using the computed rotation, scaling factor, and centroid positions. Once these transformation parameters were established, they were applied uniformly to all 468 facial landmarks in each frame. The final normalized X and Y coordinates of the keypoints across frames attributed to a 1-D array, the input features for the subsequent model. Further data cleaning also involved removing any video clips that were not exactly 1 s, as the supervised artificial intelligence (AI) model would require input data with a fixed format (size) of $28,080 \times 1$, which included the x and y coordinates of the 14,040 key points.

### Spatial-temporal-attention long-short-term-memory (STA-LSTM) model training and validation

We designed and trained a customized spatial temporal attention long short-term memory (STA-LSTM) deep learning network to detect pain levels through analyzing the dynamic facial expressions in both spatial and temporal domains (Fig. 1)[21]. Long short-term memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to capture and remember information over long sequences, which is crucial in pain detection involving time-series analysis[22]. In addition to LSTM, a spatial temporal attention (STA) mechanism was incorporated, allowing the model to focus on specific regions of interest (e.g., eye, mouth, brow) within each frame while considering the temporal evolution of these regions across multiple frames. This can enhance the model's ability to recognize facial expressions of pain that unfold over time. The input to the STA-LSTM model consisted solely of the x, y coordinates of 468 facial keypoints extracted from video data. Importantly, no predefined or handcrafted facial features (e.g., specific action units or geometric features) were used as inputs.

The model was trained on a desktop workstation running a 64-bit Windows operating system (Windows 10) with an Intel Xeon Silver 4208 processor (8 cores, 3.2 GHz), 64 GB of random-access memory (RAM), and two Nvidia RTX 2080 Ti graphics processing units (GPUs). The training and validation processes were conducted using Python 3.6, PyTorch, Mediapipe, and OpenCV. Among the 10,274 1 s clips, videos from 160 patients (7599 clips) were used for STA-LSTM training, while the remaining 40 patients' videos (2675 clips) were set aside for validation. A personalized training mechanism was also employed to further improve the accuracy of the model. The following hyperparameters were used after fine-tuning: 64 LSTM hidden layers, a learning rate of 0.025, weight decay set of 0.000001, and a batch size of 1000.

## Statistical analysis

In our previous pilot study, 26 patients were recruited with 119 videos (~ 1:5) collected that showed a pain score > 0 on the NRS[23]. We tested the proposed model using the existing University of Northern British Columbia (UNBC)-McMaster Shoulder Pain Expression Archive Database, which contains a larger dataset (> 1,000) and has an accuracy rate of > 95%[24]. By collecting videos from 200 patients in the current study, we retrieved at least 5 videos per patient with a pain score > 0, indicating that at least 1,000 videos could be used for training and validation. Patient demographics and pain outcomes were summarized using number (proportion), mean (standard deviation (SD)), or median (interquartile range (IQR)), as appropriate. Model performance was assessed using accuracy, sensitivity, recall, and F1-score (Table 1)[25]. All analyses were performed using SAS® version 9.4.

## Results

A total of 200 participants were recruited for the study. The demographic and clinical characteristics are reported in Table 2. Of note, female adult patients comprised the majority of the recruited cohort (70%) from KKH Women's services. The patients recruited from KKH (n = 100) underwent gynecological surgery, with the majority undergoing hysterectomy (n = 43), myomectomy (n = 32), and cystectomy (n = 17). At the SGH site, almost half of the patients were recruited during their pain clinic consultation (91 or 45.5%), while the remaining patients underwent steroid injection and/or radiofrequency ablation interventional pain procedures during their visit. Existing pain prior to surgery or interventional pain procedures was reported by 89 patients, with a median (IQR) score of 3 (2–6) in this group of patients with pain. Pain scores were similar after surgery or interventional pain procedures, with 92 patients reporting pain with a median (IQR) score of 3 (2–4). Post-operative analgesia was provided to KKH patients, with the majority being prescribed on paracetamol (n = 91), morphine (n = 50), etoricoxib (n = 41), mefenamic acid (n = 20), and parecoxib (n = 10).

We collected 2,008 videos (of varying duration) for the categorization of pain levels, of which 90.4% (in terms of duration) were labelled as no pain, and only 3.4 and 6.3% were labelled as significant pain and mild pain, respectively (Table 3). If such an imbalanced dataset was directly used for training the deep learning network, it could lead to a model that is biased toward the majority class (no pain), resulting in poor performance on the minority classes (mild pain and significant pain). Therefore, we adopted a balancing approach to the dataset prior to the model training. This was achieved using an adaptive stride resampling technique, where the dataset was balanced as follows: for the class of no pain, a larger stride was used to select the frames less frequently and reduce redundant data from this dominant class; while for the class of significant pain, a smaller stride (closer frame interval) was used to extract more feature vectors from the available data[26]. After balancing, the number of instances in each class became approximately equivalent, as shown in (Table 3).

By differentiating the polychromous levels of pain (no pain versus mild pain versus significant pain requiring clinical intervention), we reported optimal performance of STA-LSTM model, with the accuracy, sensitivity, recall, and F1-score all at 0.8660 (Table 4). The performance of the dichotomous levels of pain (no pain versus significant pain) was higher than for the polychromous categorization, with accuracy, sensitivity, recall, and F1-score at 0.9384, 0.9952, 0.7836, and 0.8778, respectively. Similarly, dichotomous categorization of no pain and mild pain versus significant pain yielded higher performance than polychromous categorization, having 0.9087, 0.9917, 0.8210, and 0.8983 in the accuracy, sensitivity, recall, and F1-score respectively. The area under the receiver-operating curve (ROC) for each comparison is shown in (Fig. 3). The developed STA-LSTM model was also integrated as personal computer (PC) software where patients could have their pain levels detected in

| Performance metrics | Description |
|---|---|
| True positive (TP) | The instance of correct prediction(s) as a positive class |
| True negative (TN) | The instance of correct prediction(s) as a negative class |
| False positive (FP) | The instance of incorrect prediction(s) as a positive class, with the actual class being negative |
| False negative (FN) | The instance of incorrect prediction(s) as a negative class, with the actual class being positive |
| Accuracy | The proportion of correctly predicted instance(s) (TP + TN) out of all instances (TP + TN + FP + FN) |
| Precision | The proportion of the instance(s) of correct prediction(s) as a positive class (TP) out of all predicted positive instances (TP + FP) |
| Recall | The proportion of the instance(s) of correct prediction(s) as a positive class (TP) out of all actual positive instances (TP + FN) |
| F1 | A harmonic mean of precision and recall, and is calculated as 2 × precision x recall/(precision + recall). An F1 score of 1 indicates its best in quality classification whereas 0 refers to the worst |

**Table 1.** The performance metrics and their descriptions[25].

| Parameter | Training cohort n = 160 | Validation cohort n = 40 | Overall n = 200 |
|---|---|---|---|
| Age, y | 47.4 (11.8) | 45.1 (12.8) | 46.9 (12.0) |
| Gender | | | |
| Female | 108 (67.5%) | 32 (80.0%) | 140 (70.0%) |
| Male | 52 (32.5%) | 8 (20.0%) | 60 (30.0%) |
| Race | | | |
| Chinese | 96 (60.0%) | 26 (65.0%) | 122 (61.0%) |
| Malay | 28 (17.5%) | 6 (15.0%) | 34 (17.0%) |
| Indian | 20 (12.5%) | 3 (7.5%) | 23 (11.5%) |
| Others | 16 (10.0%)[a] | 5 (12.5%)[b] | 21 (10.5%) |
| Weight, kg | 68.9 (17.0) | 70.8 (15.6) | 69.3 (16.7) |
| BMI, kg.m$^{-2}$ | 26.1 (5.4) | 27.0 (5.1) | 26.3 (5.3) |
| ASA status | | | |
| I | 42 (26.3%) | 11 (27.5%) | 53 (26.5%) |
| II | 106 (66.3%) | 26 (65.0%) | 132 (66.0%) |
| III | 12 (7.4%) | 3 (7.5%) | 15 (7.5%) |
| Purpose of visit | | | |
| Pain consultation | 72 (45.0%) | 19 (47.5%) | 91 (45.5%) |
| Gynecological surgery | 80 20 (50.0%) | 20 (50.0%) | 100 (50.0%) |
| Procedure | 7 (4.4%) | 1 (2.5%) | 8 (4.0%) |
| Pre-procedure | 1 (0.6%) | 0 | 1 (0.5%) |
| Pre-surgical/procedural pain score (at rest) | 0 (3 [0–9]) | 0 (3 [0–7.5]) | 0 (3 [0–9]) |
| Pre-surgical/procedural EQ-5D-3L health VAS score | 72.9 (16.8) | 73.0 (18.5) | 72.9 (17.1) |
| Post-surgical/procedural pain score (at rest) | 2 (3 [0–10]) | 2 (3 [0–7]) | 2 (3 [0–10]) |

**Table 2**. Demographic and clinical characteristics. Values are mean (SD), median (IQR [range]) or number (proportion). *ASA* American society of anesthesiologists, *BMI* body mass index, *VAS* visual analog scale. [a]Others included Filipino (n = 8), Indonesian (n = 4), Burmese (n = 1), Bangladeshi (n = 1), and Myanmese (n = 1). [b]Others included Filipino (n = 2), Burmese (n = 1), Eurasian (n = 1), and Pakistani (n = 1).

| | Before balancing | | | After balancing | | |
|---|---|---|---|---|---|---|
| | No. of clips | Duration of videos (HH:MM:SS) | Duration of videos (%) | No. of 1 s clips | Duration of videos (HH:MM:SS) | Duration of videos (%) |
| Significant pain | 111 | 0:17:28 | 3.4 | 3562 | 0:59:22 | 34.6 |
| Mild pain | 180 | 0:32:37 | 6.3 | 3640 | 1:00:40 | 35.4 |
| No pain | 1,717 | 7:50:13 | 90.4 | 3072 | 0:51:12 | 29.9 |

**Table 3**. Distribution of the pain videos. HH:MM:SS refers to hours:minutes:seconds.

| | No pain vs mild pain vs significant pain | No pain vs significant pain | No pain and mild pain vs significant pain |
|---|---|---|---|
| True positive | – | 413 | 1078 |
| True negative | – | 1353 | 1353 |
| False positive | – | 114 | 235 |
| False negative | – | 2 | 9 |
| Accuracy | 0.8660 | 0.9384 | 0.9087 |
| Precision | 0.8660 | 0.9952 | 0.9917 |
| Recall | 0.8660 | 0.7836 | 0.8210 |
| F1 | 0.8660 | 0.8778 | 0.8983 |

**Table 4**. The categorization of pain level, with no pain as the reference in different comparisons.
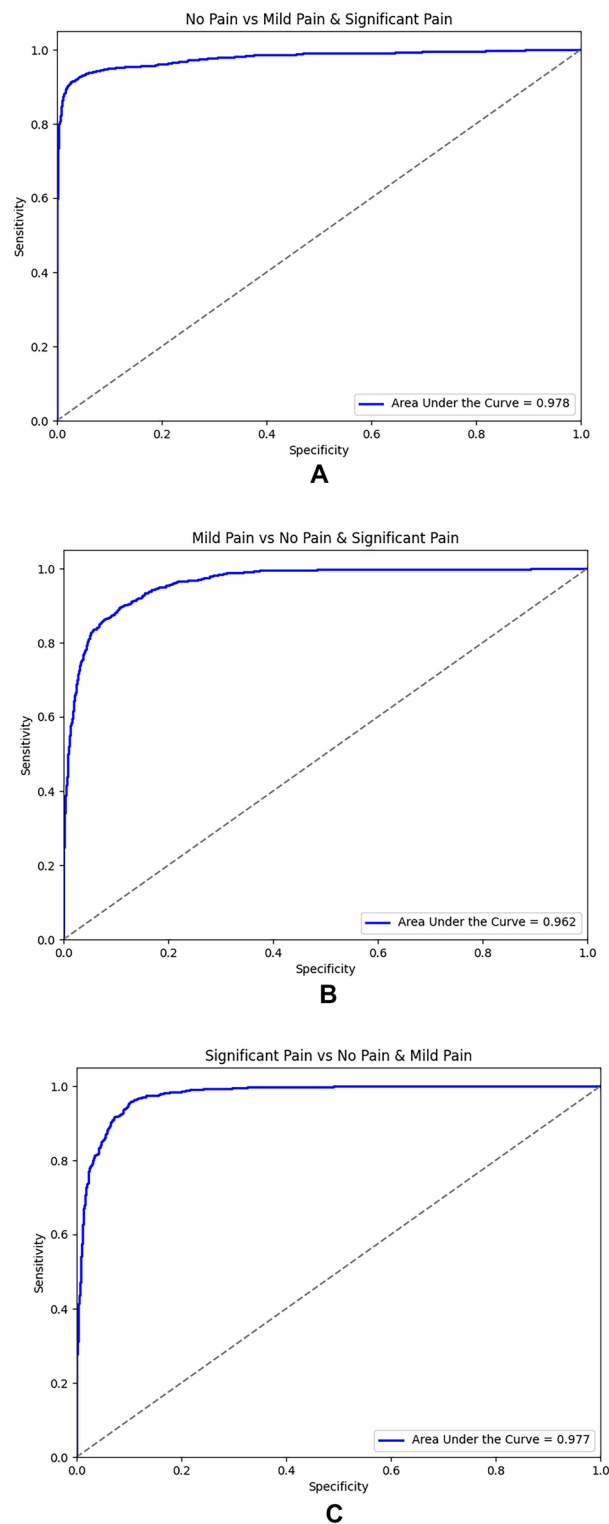
**Fig. 3**. The area under the receiver-operating curve (ROC) for (**A**) no pain versus mild pain and significant pain; (**B**) mild pain versus no pain and significant pain; and (**C**) significant pain versus no pain and mild pain.
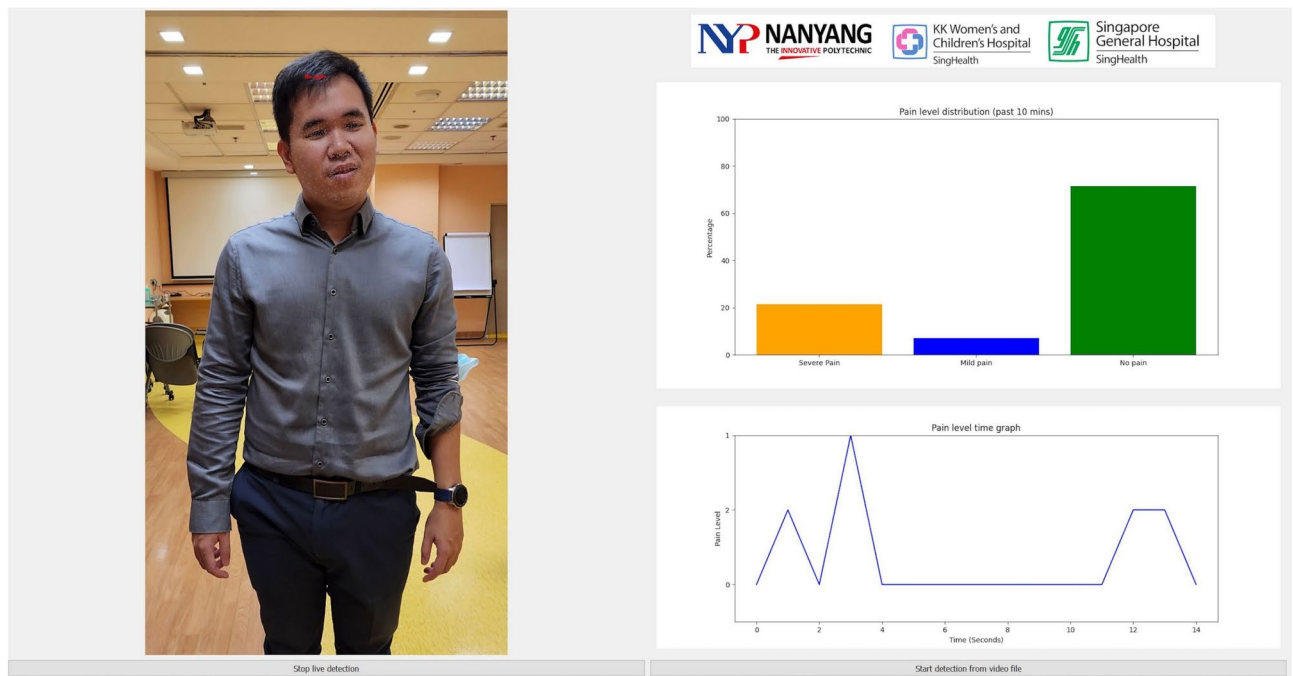
**Fig. 4**. The developed STA-LSTM model integrated into a PC software with real-time pain level distribution and time graph.

various front-facing postures (sitting, standing, and movement) in real-time manner, with pain level distribution and time graphs displayed on the same PC user interface (Fig. 4).

## Discussion

We developed a pain recognition model using a dataset from patients undergoing surgery or interventional pain procedures collected from two hospitals in Singapore. By differentiating the polychromous levels of pain (no pain, mild pain, significant pain) and training a customized STA-LSTM deep learning network with a balanced dataset, we achieved high accuracy (>85%) in the model validation.

During the data collection and analysis, several study approaches were considered. As compared with the conventional pain scoring systems such as the 0 to 10 NRS or VAS, we utilized a three-level pain categorization and defined a NRS cut-off of 4 as significant pain. We successfully addressed the uneven distribution of pain severity across the dataset; however, we did not explore the gender differences in the cohort due to the imbalanced sample size between male and female patients. Additionally, we acknowledge that pain self-reporting is considered the gold standard in clinical pain assessment due to the subjective nature of pain. However, we employed an observer rating scale to annotate the 1-s clips to mimic observer based visual assessment, which is in contrast to the conventional self-reported pain assessments in other studies[27,28]. Given that our developed model replicates the observer training process by assessing pain solely through facial expressions, it inherently aligns with observer ratings rather than subjective patient reporting, while at the same time avoids misinterpretations that could arise from labeling the entire videos based solely on patient single pain reporting. Previous report showed that the training model on assessing pain intensity based on facial expression performs better than those based on self-reporting pain[29]. This difference may be attributed to varying facial expressions and social motives (e.g., seeking social support, confronting social threat) during pain stimuli, which can lead to significant variations even across a small sample size[30,31]. Further work should address the subjective nature of pain experience by validating the model against the conventional pain scoring systems to assess its ability to discriminate between different pain levels. During the model development, we also employed an adaptive stride sampling approach to increase the number of training instances for these pain categories without artificially generating data[26]. Unlike synthetic oversampling techniques (e.g., synthetic minority over-sampling technique (SMOTE)), our method preserved natural variations in facial expressions and maintained time-series consistency by avoiding random frame shuffling and ensuring that temporal dependencies within each 1-s clip remained intact[31,32]. Most importantly, this approach reduced bias toward the majority class, enabling the model to equally prioritize different pain levels and improve accuracy in detecting mild and severe pain cases. Nevertheless, this method may also introduce over-detection of pain and increase the risk of false positives and unnecessary interventions, hence a closer monitoring is warranted during clinical applications.

To the best of our knowledge, this is the first study that utilizes STA-LSTM model in facial pain recognition. We tested several deep learning (DL) models for training, including STA-LSTM, LSTM, Convolution-Neural-Network (CNN)-LSTM and 3D CNN; and STA-LSTM exhibited the best performance among all. It has been previously reported that facial recognition is more reliable when analyzing a sequence of video frames rather than

a single frame[33]. In our study, the input to our AI model consists solely of facial keypoint coordinates rather than raw video or image data due to confidentiality constraints. CNNs are primarily designed for image analysis and hence we selected LSTM instead given the time-series nature of our data. Besides, LSTM incorporates feedback connections that can process not only single frames but also an entire video, which is ideal for learning the order dependence in sequence prediction problems[34]. We also incorporated a spatial temporal attention module into the LSTM model, as not all 468 facial keypoints are equally important for pain detection. The addition of spatial attention module allows the network to focus on pain-related keypoints (i.e., facial features that are most relevant to pain), while the temporal attention module helps the network to prioritize the most pertinent frames within the video. These enhancements improve the model's ability to capture subtle pain-related expressions more effectively compared to a basic LSTM.

Various state-of-the-art deep learning models, such as CNN and CNN-RNN architectures, have been employed in the field of pain detection. The STA-LSTM model stands out by integrating spatial and temporal attention mechanisms, allowing it to focus on the most informative features over time[35]. This approach is particularly beneficial for analyzing complex, time-dependent data, such as facial expressions associated with pain[36]. CNNs excel at extracting spatial features from static images and are commonly used to analyze individual frames for facial expressions indicative of pain. Since CNNs process each frame independently, they lack the ability to capture the temporal dynamics inherent in pain expressions and hence crucial contextual information may be missed for accurate pain assessment[37]. In view of the temporal shortcomings of standalone CNNs, hybrid models combining CNNs with RNN such as LSTM networks have been developed by extracting spatial features from each frame, followed by feeding into LSTMs to model temporal dependencies. CNN-LSTM models could outperform standalone CNNs in pain detection, reaching an accuracy of 91.2% in classifying pain types[38]. The STA-LSTM model further enhances the hybrid CNN-RNN approach by incorporating attention mechanisms that assign varying importance to different spatial features and time steps. This enables the model to focus on critical facial regions and key moments that are most indicative of pain, thereby improving detection accuracy. By dynamically adjusting its focus, STA-LSTM effectively captures subtle and transient pain expressions that other models might overlook[35,36]. Overall, the STA-LSTM model enhances the hybrid CNN-RNN approach by incorporating attention mechanisms that weigh the importance of different spatial features and time steps. This allows the model to focus on critical facial regions and specific instances that are most indicative of pain to improve detection accuracy. By dynamically adjusting its focus, STA-LSTM can effectively handle subtle and transient pain expressions that other models might overlook. Despite its strengths, STA-LSTM has certain limitations. Its complexity can lead to increased computational requirements, making real-time applications challenging without adequate resources. Additionally, while attention mechanisms enhance focus on relevant features, they can also introduce biases if not properly calibrated, potentially affecting the model's generalizability across diverse populations[35,36].

Previous studies have highlighted the use of computer vision and machine learning methods for automatic pain detection via facial expressions. However, there are recent research advances on employing DL models to recognize facial pain expressions. Rodriguez et al. introduced the combination of CNN and LSTM networks to analyze the video-level temporal dynamics in the facial features from both the UNBC-McMaster shoulder pain expression archive and cohn Kanade + facial expression databases, with the area under the curve (AUC) scores ranging from 89.6 to 93.3[39]. Similarly, Bargshady et al. developed an ensemble deep learning framework (EDLM) model that integrates three independent CNN-RNN for pain detection via facial expressions, and evaluated using UNBC-McMaster shoulder pain expression archive and multimodal intensity pain (MIntPAIN) databases[40]. Their results were promising, with an AUC of 93.7 compared to single hybrid DL models. Nevertheless, most of these studies were used existing databases from either healthy volunteers or subjects with minimal pain or evoked pain; and rarely explore the model's application beyond controlled settings, which may not be applicable in the real-life clinical scenarios. Our developed STA-LSTM model demonstrated competitive performance with an AUC of > 96.0. In this context, it is noteworthy that a high AUC does not guarantee clinical relevance nor the optimal threshold during decision making[41]. Hence, we also looked at other performance metrics including accuracy and recall that further showed that our model indeed outperformed other models from previous studies to achieve the start-of-the-art results in clinical settings[25]. Nevertheless, other metrics such as the clinical utility index, risk-benefit analysis, and model calibration, should be considered in future prospective validation to better address the clinical impact of the model in pain assessment[42]. In addition, we also acknowledge the concern regarding the potential overfitting due to the increased number of frames from the same individuals in the mild and significant pain groups. To mitigate such risk, we utilized a balanced sampling approach to ensure that the frames were not consecutively selected but were rather extracted at different time points to reduce the likelihood of overfitting to specific facial patterns from particular individuals[43]. Apart from applying dropout layers and weight regularization, we also ensured no overlapping patients between the training and validation cohorts[44,45]. Lastly, we also considered the temporal variations of facial expressions with each frame representing a unique variation of pain expression. This ensures that the model learns from diverse facial movements rather than memorizing specific identities[46].

To date, automated pain score reporting and visualization are not present in existing clinical practices. A survey of anesthetists and nursing staff found that more than 80% of the respondents indicated they would be likely or willing to use an automated pain recognition platform, and this application could benefit patients with limited communicative abilities by preventing analgesic overdose or underdose[47]. Reliability and usability were also identified as the most important factors for successful implementation[47]. The use of an automated pain assessment platform would potentially offer a more accurate, robust, and timely pain assessment and improve reporting, compliance, and monitoring of pain management in various clinical settings, particularly for non-communicative populations. For instance, this platform could enable the caregivers of elderly or non-communicative patients (stroke, cognitive deficit, dementia etc.) to facilitate pain assessment in outpatient

settings. Similarly, future telehealth applications could utilize facial recognition to diagnose pain-related conditions, including immobility or dysfunction which may not be evident during a telehealth consultation. This may in turn improve the pain management and care, including but not limited to better prognosis for recovery, accurate titration of pain medication, reduced overall opioid analgesics consumption and opioid-associated adverse effects. Subsequently, these will translate to improved pain experience following discharge from hospital, and subsequently reduce psychological and social sequelae of both patients and caregivers. Notably, this AI-based technology may also raise challenges regarding accountability in cases of pain misclassification that lead to under- or overtreatment. Currently, AI-driven software is categorized as software as a medical device (SaMD) under the international medical device regulators forum (IMDRF) and the U.S. Food and Drug Administration (FDA); and must comply with regulations such as the EU medical device regulation (MDR) 2017/745[48,49]. Subject to different liabilities under malpractice, product liability, and negligence theories, multiple stakeholders—including developers, healthcare institutions, and attending clinicians—should collaborate to identify systemic flaws and validate outputs through clinical judgment and training before implementation[50]. Nonetheless, it is important to note that such AI-based technology should complement clinical decision-making rather than replace human judgment. It should provide explainable reasoning for the generated pain scores, with the final decision-making authority retained by clinicians, including overriding AI recommendations when necessary[51].

At this stage, we have successfully integrated the developed STA-LSTM model as a PC software, with real-time pain level distribution displayed in the user interface (Fig. 4). However, we acknowledge that the model was developed based on the data from individuals who would not the typical primary candidates of the proposed model as they were able to self-report pain. Conversely, the model should be further validated in non-verbal patients especially those with neurological conditions that may affect the facial expression and the subsequent idiosyncratic pain expressions that may affect the pain predictions. Future work will continue to prospectively validate the model in diverse clinical and community settings especially in the non-verbal populations, such as Parkinson's disease, cerebral palsy, and dementia[52]. We are also liaising with the relevant parties on the local and international regulatory and cybersecurity guidelines, risk management and mitigation before driving the development for integration into existing clinical workflows. This includes corporate electronic medical record (EMR) applications to convey the pain results intuitively to clinicians alongside other clinical data, and evaluate its impact upon clinical decision making and patient care. We note that many official EMR systems impose security and operational protocols that do not allow feature integration from third-party developer systems[53]. Hence, we are also exploring ways to integrate the model into our customized mobile application either as a commercial product or medical device that complies with international standards (e.g., European Committee for Standardization (CEN)-International organization for standardization (ISO)/technical specification (TS) 82304-2), which could serve as a community-based healthcare platform to enhance accessibility and convenience for patients in outpatient settings[54]. Lastly, we are also aware of the sensitivity and privacy issues brought by the patient facial features, and the possible patient discomfort and distrust on the automated pain assessment. The keypoint extraction will continue to be implemented as real-time encryption and anonymization for data protection, with patient consent and education process further refined to allow opting in/out options and non-intrusiveness during recording.

We acknowledge several limitations in this study. Our study utilized dataset from a predominantly multi-ethnic Asian population residing in Singapore, consisting of individuals with either chronic pain issues, or undergoing surgical procedures. It was previously reported that Asians had a lower pain threshold and perceptions as compared to other ethnicities, thus the generalizability of the findings should be cautiously considered[55]. Notably, almost half of our study cohort were attended for issues relating to chronic pain, which may significantly alter pain perception and result in differences in facial pain expression and reporting[56]. A more diverse population representing different sociocultural settings (e.g., age, gender, ethnicities, clinical profile) could help enhance the generalizability of the findings. AI model development often involves large datasets, however the dataset collected in this study is limited by the small sample size and number of videos representing significant and mild pain, which may potentially reduce the sensitivity of these groups of pain. This limitation is attributed to the ethical study design, where patients could decline to perform an action (and associated pain rating) at their discretion, which resulted in the loss of video(s) and a reduced number of clips for training. Expanding similar studies in patients undergoing surgeries that yield more intense nociception (e.g., thoracic, major abdominal surgeries), with different pain conditions, or including post-operative assessment through actions that may invoke more pain (e.g., incentive spirometry, cough), may help improve the dataset for significant pain. We also discussed the use of observer rating in this study, which could introduce potential limitation on the model training based on observers' interpretations rather than the patients' actual pain experiences.

Additionally, we were unable to perform subgroup analyses based on gender, surgical type, and chronic pain status due to the limited sample size. Among the validation set, there were 116 videos (4.3%) having pain mismatches, i.e., pain levels not at the adjacent level (e.g., significant versus no pain, or vice versa). Nevertheless, we did not trace the raw videos to investigate the cause due to the nature of facial keypoint extraction, nor perform any statistical inference testing between the subsamples as the individual-level identifiers were not retained during the processing of anonymized data. We employed such random selection process to ensure an even distribution of the pain levels and clinical conditions, but tagging would be necessary in the future to regain such traceability for further inference testing. Additional analysis on the breakdown of facial characteristics that correspond to the pain levels, as well as other data (e.g., body pose, substance and analgesia used, psychological measures, pain perception and thresholds) could improve the performance of the model. Further data collection on these factors will be useful to refine the model complexity for future applications, such as physiotherapy and telehealth settings. AI-generated models are commonly known to struggle with generalization, as the developed model may inherit biases from the training data, leading to unfair or inequitable outcomes[57]. Besides, the data distribution during training may differ from that of other demographic groups or contexts, which can result in a

drop in performance. Thus, approaches such as fairness-aware training and domain adaptation may be required to improve our model and mitigate such issues[58].

## Conclusions

We developed a model that could serve as an automated pain assessment platform with high accuracy in patients undergoing surgery or interventional pain procedures. Further validation and refinement are warranted to extend the application of this model in both inpatient and outpatient healthcare settings, enabling healthcare professionals and caregivers to perform pain assessments effectively.

## Data availability

The datasets generated and/or analyzed during this study are not publicly available due to institutional policy on data confidentiality but are available from the corresponding author on reasonable request.

## References

1. LeResche, L. Facial expression in pain: A study of candid photographs. *J. Nonverbal. Behav.* **7**, 46–56. https://doi.org/10.1007/BF01001777 (1982).
2. Skrobik, Y. & Flood, P. Pain, analgesic effectiveness, and long-term opioid dependency. In *Post-Intensive Care Syndrome* (eds Preiser, J. C. et al.) 213–222 (Springer, 2020).
3. Thong, I. S., Jensen, M. P., Miró, J. & Tan, G. The validity of pain intensity measures: what do the NRS, VAS, VRS, and FPS-R measure?. *Scand. J. Pain* **18**, 99–107. https://doi.org/10.1515/sjpain-2018-0012 (2018).
4. Craig, K. C., Prkachin, K. M. & Grunau, R. E. The facial expression of pain. In *Handbook of Pain Assessment* (eds Turk, D. C. & Melzack, R.) (Guilford Press, 2001).
5. Manocha, S. & Taneja, N. Assessment of paediatric pain: a critical review. *J. Basic Clin. Physiol. Pharmacol.* **27**, 323–331. https://doi.org/10.1515/jbcpp-2015-0041 (2016).
6. Kunz, M., Meixner, D. & Lautenbacher, S. Facial muscle movements encoding pain—A systematic review. *Pain* **160**, 535–549. https://doi.org/10.1097/j.pain.0000000000001424 (2019).
7. Dildine, T. C. & Atlas, L. Y. The need for diversity in research on facial expressions of pain. *Pain* **160**, 1901–1902. https://doi.org/10.1097/j.pain.0000000000001593 (2019).
8. Chen, Z., Ansari, R. & Wilkie, D. J. Automated pain detection from facial expressions using FACS: A review. *arXiv* https://doi.org/10.48550/arXiv.1811.07988 (2018).
9. Atee, M., Hoti, K., Parsons, R. & Hughes, J. D. Pain assessment in dementia-evaluation of a point-of care technological solution. *J. Alzheimers Dis.* **60**, 137–150. https://doi.org/10.3233/JAD-170375 (2017).
10. Hoti, K., Atee, M. & Hughes, J. D. Clinimetric properties of the electronic Pain assessment tool (ePAT) for aged-care residents with moderate to severe dementia. *J. Pain Res.* **11**, 1037–1044. https://doi.org/10.2147/JPR.S158793 (2018).
11. Hassan, T. et al. Automatic detection of pain from facial expressions: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1815–1831. https://doi.org/10.1109/TPAMI.2019.2958341 (2019).
12. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* **131**, 211–219. https://doi.org/10.1161/CIRCULATIONAHA.114.014508 (2015).
13. EuroQoL Group. EuroQol—A new facility for the measurement of health-related quality of life. *Health Policy* **16**, 199–208. https://doi.org/10.1016/0168-8510(90)90421-9 (1990).
14. Gélinas, C., Fillion, L., Puntillo, K. A., Viens, C. & Fortier, M. Validation of the critical-care pain observation tool in adult patients. *Am. J. Crit. Care* **15**, 420–427. https://doi.org/10.4037/ajcc2006.15.4.420 (2006).
15. Wu, C. L. et al. Deep learning-based pain classifier based on the facial expression in critically ill patients. *Front. Med.* **9**, 851690. https://doi.org/10.3389/fmed.2022.851690 (2022).
16. Gerbershagen, H. J., Rothaug, J., Kalkman, C. J. & Meissner, W. Determination of moderate-to-severe postoperative pain on the numeric rating scale: a cut-off point analysis applying four different methods. *Brit. J. Anaesth.* **107**, 619–626. https://doi.org/10.1093/bja/aer195 (2011).
17. Kartynnik, Y., Ablavatski, A., Grishchenko, I. & Grundmann, M. Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv* https://doi.org/10.48550/arXiv.1907.06724 (2019).
18. Lugaresi, C. et al. Mediapipe: A framework for building perception pipelines. *arXiv* https://doi.org/10.48550/arXiv.1906.08172 (2019).
19. Wu, Y., Gou, C. & Ji, Q. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. *arXiv* https://doi.org/10.48550/arXiv.1709.08130 (2017).
20. Levinson, J. et al. An analysis of svd for deep rotation estimation. *Adv. Neural Inf. Process. Syst.* **33**, 22554–22565 (2020).
21. Ding, Y., Zhu, Y., Wu, Y., Jun, F. & Cheng, Z. Spatio-temporal attention LSTM model for flood forecasting. In *2019 International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* 458–465 (2019)
22. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **12**, 2451–2471. https://doi.org/10.1162/089976600300015015 (2000).
23. Lee, M. et al. Pain intensity estimation from mobile video using 2D and 3D facial keypoints. *arXiv* https://doi.org/10.48550/arXiv.2006.12246 (2020).
24. Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E. & Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* 57–64 https://doi.org/10.1109/FG.2011.5771462 (2011).
25. Powers, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* https://doi.org/10.48550/arXiv.2010.16061 (2020).
26. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969 (2008).
27. Fontaine, D. et al. Artificial intelligence to evaluate postoperative pain based on facial expression recognition. *Eur. J. Pain* **26**, 1282–1291. https://doi.org/10.1002/ejp.1948 (2022).
28. Tsai, F. S., Hsu, Y. L., Chen, W. C., Weng, Y. M., Ng, C. J., Lee, C. C. Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions. In *Interspeech* 92–96 https://doi.org/10.21437/Interspeech.2016-408 (2016).

29. Othman, E. et al. Automatic vs human recognition of pain intensity from facial expression on the X-ITE pain database. *Sensors* **21**, 3273. https://doi.org/10.3390/s21093273 (2021).
30. Kunz, M., Prkachin, K. & Lautenbacher, S. Smiling in pain: Explorations of its social motives. *Pain Res. Treat.* **128093**, 1–8. https://doi.org/10.1155/2013/128093 (2013).
31. Karos, K., Meulders, A., Goubert, L. & Vlaeyen, J. W. Hide your pain: social threat increases pain reports and aggression, but reduces facial pain expression and empathy. *J. Pain* **21**, 334–346. https://doi.org/10.1016/j.jpain.2019.06.014 (2020).
32. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953 (2002).
33. Bassili, J. N. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *J. Pers. Soc. Psychol.* **37**, 2049. https://doi.org/10.1037/0022-3514.37.11.2049 (1979).
34. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270. https://doi.org/10.1162/neco_a_01199 (2019).
35. Song, S., Lan, C., Xing, J., Zeng, W. & Liu, J. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process.* **27**, 3459–3471. https://doi.org/10.1109/TIP.2018.2818328 (2018).
36. Ridouan, A., Bohi, A. & Mourchid, Y. Improving pain classification using spatio-temporal deep learning approaches with facial expressions. *arXiv* https://doi.org/10.48550/arXiv.2501.06787 (2025).
37. Hans, A. S. & Rao, S. A CNN-LSTM based deep neural networks for facial emotion detection in videos. *Int. J. Adv. Signal Image Sci.* **7**, 11–20. https://doi.org/10.29284/ijasis.7.1.2021.11-20 (2021).
38. Fernandez Rojas, R., Joseph, C., Bargshady, G. & Ou, K. L. Empirical comparison of deep learning models for fNIRS pain decoding. *Front. Neuroinform.* **18**, 1320189. https://doi.org/10.3389/fninf.2024.1320189 (2024).
39. Rodriguez, P. et al. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE Trans. Cybern.* **52**, 3314–3324. https://doi.org/10.1109/TCYB.2017.2662199 (2022).
40. Bargshady, G. et al. Ensemble neural network approach detecting pain intensity from facial expressions. *Artif. Intell. Med.* **109**, 101954. https://doi.org/10.1016/j.artmed.2020.101954 (2020).
41. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making.* **26**, 565–574. https://doi.org/10.1177/0272989X06295361 (2006).
42. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2 (2010).
43. Liu, H., et al. A novel deep framework for change detection of multi-source heterogeneous images. In *2019 International Conference on Data Mining Workshops (ICDMW)* (pp. 165–171). https://doi.org/10.1109/ICDMW.2019.00034 (2019).
44. Zhang, Z. et al. Efficient and generalizable cross-patient epileptic seizure detection through a spiking neural network. *Front. Neurosci.* **17**, 1303564. https://doi.org/10.3389/fnins.2023.1303564 (2024).
45. Garbin, C., Zhu, X. & Marques, O. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimed. Tool Appl.* **79**, 12777–12815. https://doi.org/10.1007/s11042-019-08453-9 (2020).
46. Szczapa, B. et al. Automatic estimation of self-reported pain by interpretable representations of motion dynamics. *Proc. IAPR Int. Conf. Pattern Recogn.* https://doi.org/10.1109/icpr48806.2021.9412292 (2021).
47. Walter, S. et al. What about automated pain recognition for routine clinical use? A survey of physicians and nursing staff on expectations, requirements, and acceptance. *Front. Med.* **7**, 566278. https://doi.org/10.3389/fmed.2020.566278 (2020).
48. US Food and Drug Administration. Software as a medical device (SaMD). (accessed 4 Mar 2025); https://www.fda.gov/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/default.htm
49. The European Parliament and the Council of the European Union. Regulation (EU) 2017/745 of the European parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. (accessed 4 Mar 2025); https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0745
50. Maliha, G., Gerke, S., Cohen, I. G. & Parikh, R. B. Artificial intelligence and liability in medicine: balancing safety and innovation. *Milbank Quart.* **99**, 629. https://doi.org/10.1111/1468-0009.12504 (2021).
51. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intel.* **1**, 206–215. https://doi.org/10.1038/s42256-019-0048-x (2019).
52. Sabater-Gárriz, Á. et al. Automated facial recognition system using deep learning for pain assessment in adults with cerebral palsy. *Digit. Health* **10**, 20552076241259664. https://doi.org/10.1177/20552076241259664 (2024).
53. Mergel, I. The long way from government open data to mobile health apps: Overcoming institutional barriers in the US federal government. *JMIR Mhealth Uhealth* **2**, e3694. https://doi.org/10.2196/mhealth.3694 (2014).
54. Hoogendoorn, P. et al. What makes a quality health app—Developing a global research-based health app quality assessment framework for CEN-ISO/TS 82304-2: delphi study. *JMIR Form. Res.* **7**, e43905. https://doi.org/10.2196/43905 (2023).
55. Watson, P. J., Latif, R. K. & Rowbotham, D. J. Ethnic differences in thermal pain responses: a comparison of South Asian and White British healthy males. *Pain* **118**, 194–200. https://doi.org/10.1016/j.pain.2005.08.010 (2005).
56. Vachon-Presseau, E. et al. Multiple faces of pain: effects of chronic pain on the brain regulation of facial expression. *Pain* **157**, 1819–1830. https://doi.org/10.1097/j.pain.0000000000000587 (2016).
57. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* **2**, 100347. https://doi.org/10.1016/j.patter.2021.100347 (2021).
58. Kouw, W. M. & Loog, M. A review of domain adaptation without target labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 766–785. https://doi.org/10.48550/arXiv.1901.05335 (2019).

## Acknowledgements

## Author contributions

Chin Wen Tan: Conceptualization, methodology, investigation, project administration, writing—original draft. Tiehua Du: Conceptualization, methodology, software, formal analysis, writing—review and editing. Jing Chun Teo: Methodology, investigation, software, data curation, writing—review and editing. Diana Xin Hui Chan: Conceptualization, methodology, investigation, project administration, writing—review and editing. Wai Ming Kong: methodology, software, formal analysis, supervision, writing—review and editing. Ban Leong Sng: Conceptualization, methodology, investigation, writing—review and editing, supervision, project administration, funding acquisition.

### Funding

### Declarations

### Competing interests

The authors declare no competing interests.

### Ethical approval and consent to participate

The study received approval from the SingHealth Centralized Institutional Review Board (Reference Number: 2019/2293) on 26 Apr 2019 and was registered on Clinicaltrials.gov (NCT04011189) on 8 Jul 2019. Written informed consent was obtained from all participants and the study is conducted in accordance with the Declaration of Helsinki.

### Consent for publication

The facial features in Figs. 2 and 4 belong to the co-authors, Ban Leong Sng and Jing Chun Teo, who have consented to their use in this manuscript.

### Additional information

**Correspondence** and requests for materials should be addressed to C.W.T.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note**  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.