Pain Scores Estimation using Surgical Pleth Index and Long Short-Term Memory Neural Networks

Omar M. T. Abdel Deen[1], Wei-Horng Jean[1,2], Shou-Zen Fan[3,4], Maysam F. Abbod[5], and Jiann-Shing Shieh[1],*
[1] Department of Mechanical Engineering, Yuan Ze University, Taoyuan 320, Taiwan
(Tel: +886-3-466-8800 ext. 2250; E-mail: jsshieh@saturn.yzu.edu.tw; omartalab40@gmail.com)
[2] Department of Anesthesiology, Far Eastern Memoria70Hospital, Banqiao District, New Taipei City 220, Taiwan (Tel: +886-2-8966-8800 ext. 1614; Email: dtpc25@gmail.com)
[3] Department of Anesthesiology, College of Medicine, National Taiwan University, Taipei 100, Taiwan (Tel: +886-2-23123456 ext. 265523; Email: shouzen@gmail.com)
[4] Department of Anesthesiology, En Chu Kong Hospital, New Taipei City 237, Taiwan (Tel: +886-2-26723456; Email: shouzen@gmail.com)
[5] Department of Electronics and Electrical Engineering, Brunel University London, London UB8 3PH, UK (Tel: +44 (0)1895 267061; Email: Maysam.Abbod@brunel.ac.uk)

Abstract: Pain monitoring is crucial to provide proper healthcare for patients during general anesthesia (GA). In this study, photoplethysmographic waveform amplitude (PPGA), heartbeat interval (HBI), and surgical stress index (SSI) are utilized for predicting pain scores during GA based on expert medical doctors' assessments (EMDA). Time series features are fed into different long short-term memory (LSTM) models, with different hyperparameters. The models' performance is evaluated using mean absolute error (MAE), standard deviation (SD), and correlation (Corr). Three different models are used, the first model resulted in 6.9271±1.913, 9.4635±2.456, and 0.5955 0.069 for an overall MAE, SD, and Corr, respectively. The second model resulted in 3.418±0.715, 3.847±0.557, and 0.634±0.068 for an overall MAE, SD, and Corr, respectively. In contrast, the third model resulted in 3.4009±0.648, 3.909±0.548, and 0.6197±0.0625 for an overall MAE, SD, and Corr, respectively. The second model is selected as the best model based on its performance and applied 5-folds cross-validation for verification. Statistical results are quite similar: 4.722±0.742, 3.922±0.672, and 0.597±0.053 for MAE, SD, and Corr, respectively. In conclusion, the SSI effectively predicted pain score based on EMDA, not only good evaluation performance, but the trend of EMDA is replicated, which can be interpreted as relation between SSI and EMDA, however, further improvements on data consistency are also needed to validate the results and obtain better performance. Furthermore, the usage of further signal features could be considered along with SSI.

Keywords: Long short-term memory networks, Surgical stress index (SSI), Surgical Pleth Index (SPI), Heartbeat interval, Photoplethysmographic waveform amplitude, Pain score.

1 INTRODUCTION

Thousands of surgeries are performed daily worldwide, making monitoring the depth of anesthesia (DoA) a clinical routine. Pain monitoring during General Anesthesia (GA) is crucial to provide appropriate patient care and to avoid an overdose or an underdose of analgesia. GA is achieved using anesthetic drugs, leading to loss of consciousness, immobility, and analgesia in surgical operations [1]. Nociception is known as the central nervous system's ability to react to painful stimuli, whereas the ability to prevent these reactions is called antinociception [2].

Surgical Stress Index (SSI) has been developed to evaluate surgical stress during GA [3]. It is derived from normalized Heartbeat Iintervals (HBI_{norm}) and norlamised photoplethysmographic ($PPGA_{norm}$) time series ranging from 0 to 100. The algorithm is developed based on the correlation of a set of parameters, including the HBI_{norm}, $PPGA_{norm}$, noninvasive blood pressure (NIBP), pulse transit time (PTT), and other parameters, as reported in [3]. These parameters are combined with the components of another parameter that represents total surgical stress (TSS) (i.e., the actual level of stress during surgery). TSS is represented by two components: the level of stimulation and the predicted remifentanil Ce_{remi} level. Another criterion for the algorithm parameters is the parameter that explained the TSS. Based on these evaluations, many parameters showed good correlation with TSS components, and the best candidates were $PPGA_{norm}$ and HBI_{norm}.

Many studies have investigated the effects of SSI on nociception, analgesia, and antinociception during surgeries [4][5][6][7][8]. SSI has shown effective results in predicting nociception–antinociception balance, reducing opioid consumption [5–8], and detecting nociceptive stimuli [4]. Artificial intelligence is widely used in the medical field. Chowdhury et al. [9] used deep learning models to predict DoA. They obtained 86% accuracy in a ten-layer model and found that electrocardiogram (ECG) signals are more valuable than photoplethysmogram (PPG) signals for predicting DoA. The use of LSTM for pain score prediction is introduced in [10], wherein the authors used the analgesia nociception index (ANI) [11, 12] to predict pain levels using LSTM and multilayer perceptron (MLP). The achieved MAE ± SD values are 2.490 ± 0.522 and 2.633 ± 0.542 for MLP and LSTM, respectively.

The use of gold standard measures during GA has been reported in some studies [13, 14, 15]. These measures are

valuable indicators of expertise among medical doctors, enabling them to evaluate anesthesia measurements based on their experience. The acquisition of gold standards can be accomplished through various approaches. In practical scenarios, doctors' experiences can play a significant role. Thus, a weighted average could be used as the gold standard, considering doctors' expertise. Conversely, in certain situations, labels provided by less experienced doctors may be excluded if they deviate significantly from those of other doctors.

SSI is an objective measure that guides anesthesia in practice. HBI and PPGA depend on the balance between sympathetic and parasympathetic reactions; however, other factors are known to affect those reactions, including chronic high arterial pressure and antihypertensive drugs [18]. In contrast, EMDA is a subjective assessment that depends on doctors' experience in analyzing clinical measurements related to anesthesia, such as heart rate (HR) and blood pressure (BP). Thus, the hypothesis of this study is based on the effective relationship observed between SSI and various anesthesia measurements, as well as the evaluation of EMDA considering factors such as doctors' experience, PPG signal characteristics, and HR, among other relevant clinical factors. The objective of this study is to evaluate the performance of LSTM in predicting pain scores during GA using SSI and EMDA.

2 METHODS

2.1 Data Collection

This work received approval from the institutional review board of the National Taiwan University Hospital (NTUH) and informed consent is obtained from all participants involved in the study. In 2015, 142 patients were undergoing surgical operations at NTUH. ECG (512 Hz) and PPG (128 Hz) waveforms were collected using an IntelliVue MP60 physiological signal monitoring device. During signal analysis, it was noticed that some signals were corrupted or sometimes the PPG waveform was missed. In both cases, patient data were excluded (i.e., 52 patients' data); thus, 90 patients' data were used.

EMDA was measured separately after the operations were completed. The medical team recorded all needed information during the operation (e.g., HR, BP, and other clinical factors), after which expert anesthesiologists analyzed these factors and assessed the pain levels of each patient, ranging from 0 to 100. Each patient's assessment was checked using the interclass correlation (ICC) test to validate the reliability of the doctors' labels. A good ICC value ranges from 0.6 to 0.74, with values < 0.4 are considered poor [19].

2.2 Data Synchronization

The specific time at which each doctor began assessing pain scores may have differed, but they all followed the same process. Therefore, we matched the durations for all the doctors' data by comparing the durations for each patient. For differences in the beginning and ending times, these offsets were excluded, and the data durations were matched. The ECG and PPG signals were also matched, depending on the matched pain assessments.

For each patient's assessment, we found the start time t_start_i and end time t_end_i. We defined a set T_start = {t_start_1, t_start_2, ..., t_start_N} to be the set of start times for all N assessments of the patient, and we defined a set T_end = {t_end_1, t_end_2, ..., t_end_N} to be the set of end times for all N assessments. We then selected the lower limit for the synchronization algorithm as the maximum value of t_start_i, i.e., L = max(T_start). This is the latest start time among all assessments and ensures that the algorithm does not use any data points prior to this time. All data points before L were removed.

Similarly, we selected the upper limit for the algorithm as the minimum value of t_end_i, i.e., U = min(T_end). This is the earliest end time among all assessments and ensures that the algorithm does not use any data points from after this time. All data points after U were removed. The synchronized data between the lower limit L and the upper limit U were then used for further analysis. In mathematical notation:

$$T\_start = \{t\_start\_1, t\_start\_2, ..., t\_start\_N\} \quad (1)$$

$$T\_end = \{t\_end\_1, t\_end\_2, ..., t\_end\_N\} \quad (2)$$

$$L = \max(T\_start) \quad (3)$$

$$U = \min(T_{end}) \quad (4)$$

Then we performed the same procedures on ECG and PPG signals. Let L and U be the lower and upper limits, respectively, determined in the previous step. Then, the synchronized ECG and PPG signals, denoted as ECG_sync(t) and PPG_sync(t), respectively, are given by:

$$ECG\_sync(t) = \{ECG(t), \text{if } L \le t \le U; 0, \text{otherwise}\} \quad (5)$$

$$PPG\_sync(t) = \{PPG(t), \text{if } L \le t \le U; 0, \text{otherwise}\} \quad (6)$$

The synchronized signals only retain the data points within the time interval [L, U]. All data points before L and after U are set to zero (i.e., dropped). This process ensured that the ECG and PPG signals were aligned with the synchronized pain score data for further analysis. Fig. 1 shows pain score readings for a patient from 4 medical doctors with their durations, and Fig. 2 shows the duration of all medical doctors' assessments after data synchronization, with the average of the 4 assessments.

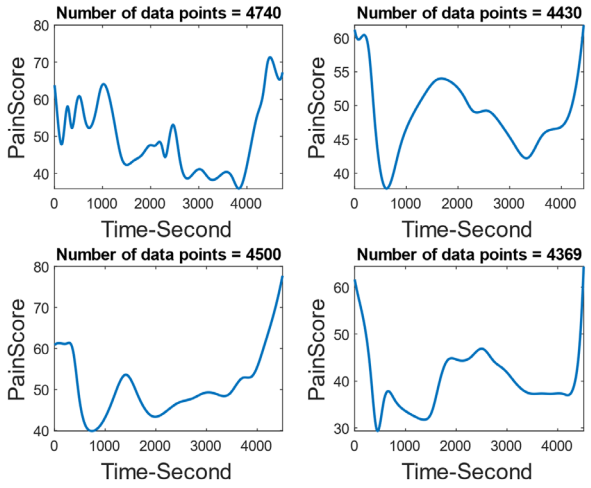

Fig. 1 Pain assessment with the time duration from each doctor before matching the time.

Agreement between the assessments is checked after synchronization was completed. The ICC test provided a statistical description of the consistency or agreement between the groups. In this study, 71 cases had good to excellent ICC measures, 18 cases had moderate ICC measures, and 1 case is poor. As shown in Fig. 3, since most cases had a score > 0.6, and moderate or poor scores do not necessarily indicate no correlation, we have used all the data for further analysis.

2.3 Signal Filtering and SSI Calculation

In this study, moving average filter (MAF) is used to smooth the signal and then applied a bandpass filter to eliminate high- and low-frequency noise. Wide MAF provides the trend of a signal, whereas narrow MAF smooths the signal. The window size is set to 1.5 sec. The MAF formula is as follows:

$$y[n]=\frac{1}{N}\sum_{i=0}^{N-1} x(n-i) \quad (7)$$

where N is the window size, $y[n]$ is the output signal and $x(n - i)$ is the previous input from the input signal. Following signal smoothing, a bandpass filter was applied to eliminate high and low frequencies. A Chebyshev II 5th order filter is used for its ability to improve the quality of PPG signals [16]. The assigned cutoff frequencies were 0.5 Hz to 8 Hz.

The HBI time series is detected as time duration between two consecutive pulse peaks in PPG signals, and R peaks in ECG signals. The PPGA is detected as the amplitude of the PPG signal. Given the importance of capturing the trend of SSI rather than short-term fluctuations, as the latter could be influenced by heart arrhythmia or artifacts without affecting the long-term parameter [17], the PPGA and HBI series are lowpass filtered using MAF with a window size of 60 seconds, to avoid losing the trends and to obtain a smooth signal that helps predict pain scores for further analysis.

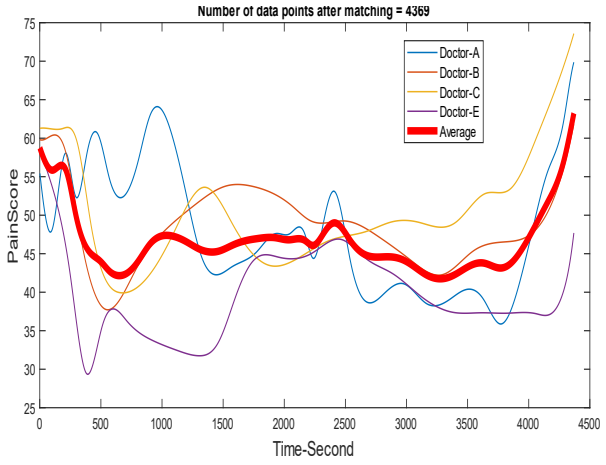

Fig. 2 Pain assessment from each doctor after matching the durations.

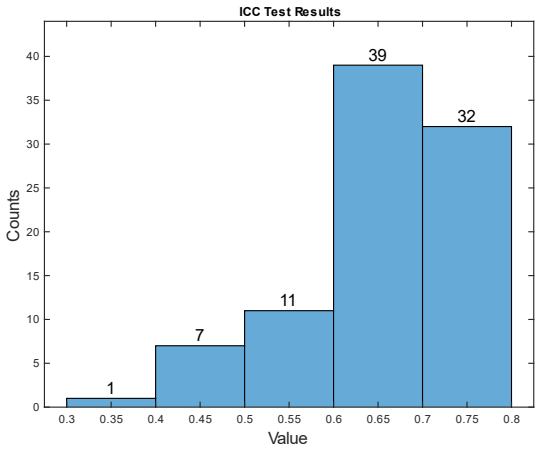

Fig. 3 ICC score proportions.

SSI calculations are performed using a combination of weighted normalized PPGA and HBI series. Normalization is required to reduce interpatient variability [3]. Histogram normalization uses a combination of two distributions: a group distribution and an individual distribution. A combination of the distributions will be used to map the values of the original HBI and PPGA.

This process is performed using MATLAB. First, all patients' PPGA and HBI data are stored in two groups,then, every patient's data is processed individually. The PPGA and HBI values vary widely between different patients, making it difficult to compare and analyze the results accurately. To address this issue, the PPGA and HBI values are normalized using a combination of two distributions: a group distribution (HBI_{group} and $PPGA_{group}$) and an individual distribution ($HBI_{individual}$ and $PPGA_{individual}$). The group distribution represents the overall distribution of PPGA and HBI values for all patients in the study, while the individual distribution represents the distribution of PPGA and HBI values for each individual patient. Weights of 0.3 and 0.7 for group and individual data are designed to reflect long-term changes in a patient's cardiovascular state

in a single numeric scale with no interpatient variability [3]. Thus, the mean (μ) and SD are calculated for the group distributions and each individual distribution, and then the mean and SD for each dataset were weighted, 0.7 for individual patients, and 0.3 for the groups as follows.

$$g\_\mu_{weighted} = 0.3 \times group\_\mu \quad (11)$$

$$I\_\mu_{weighted} = 0.7 \times individual\_\mu \quad (12)$$

$$g\_SD_{weighted} = 0.3 \times group\_SD \quad (13)$$

$$I\_SD_{weighted} = 0.7 \times individual\_SD \quad (14)$$

where $g\_\mu_{weighted}$, $I\_\mu_{weighted}$, $g\_SD_{weighted}$, and $I\_SD_{weighted}$ are group weighted-mean, individual weighted-mean, group weighted-SD, and individual weighted-SD, respectively.

Then the distributions are combined by taking the summation of the weighted μ and SD, and the cumulative distribution for the combined data is estimated to map the original HBI and PPGA time series.

$$SD_{Combined} = \sqrt{(g\_SD_{weighted})^2 + (I\_SD_{weighted})^2} \quad (15)$$

$$\mu_{Combined} = g\_\mu_{weighted} + I\_\mu_{weighted} \quad (16)$$

The SSI is then calculated based on equation (17), where $PPGA_{norm}$ is weighted by 0.7, and HBI_{norm} by 0.3, indicating that $PPGA_{norm}$ has more contribution to the algorithm. The weights are determined using the least-squares fit with TSS estimate, where the factors are in ratio 1:0.44 for $PPGA_{norm}$ and HBI_{norm} [3]. The selection process depended on the variables that explained TSS. First, the model used $PPGA_{norm}$ data, and as more variables are added, the unexplained variability decreased. However, a model with more than two variables did not have a significant improvement, hence, the SSI equation is constructed as follows:

$$SSI = 100 - (0.7 \times PPGA_{norm} - 0.3 \times HBI_{norm}) \quad (17)$$

Originally, SSI is calculated using HBI_{norm} from ECG signals, however, many signals had low quality, which resulted in inaccurate measures. Therefore, the HBI is calculated from PPG signals, where in this case the index is still representing the same quantity i.e., surgical stress, but with different name i.e., Surgical Pleth Index (SPI). The SPI (GE Healthcare, Helsinki, Finland) is a dimensionless score which is based on the photoplethysmographic analysis of the pulse wave and the heartbeat interval [18-19]. Therefore, from here on, in this study, the surgical stress index is referred as SPI. Hence, equation (17) becomes:

$$SPI = 100 - (0.7 \times PPGA_{norm} - 0.3 \times HBI_{norm}) \quad (18)$$

SPI findings are compared with EMDA to better understand our input and output, and to find any potential correlation between the trends of the two variables. As shown in Fig. 4, EMDA has higher SD values, which emphasizes the importance of considering the variability and distribution of data when interpreting the relationship between SPI and EMDA. Despite the higher variability in EMDA, the observed correlation suggests a potential connection between the two variables. Figure 4 shows a minimum correlation of 0.55, indicating that it has a similar trend, as SPI high values stand for high pain, and low SPI values stand for low pain.

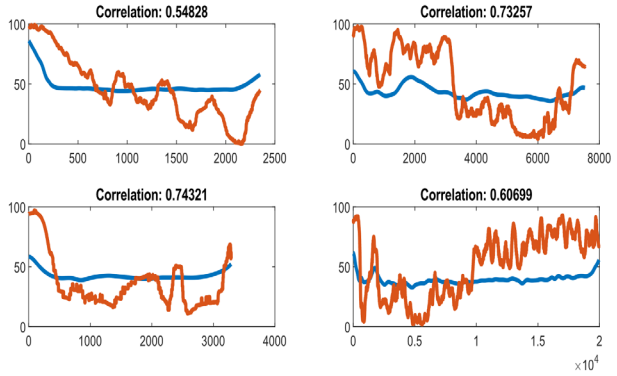

Fig. 4 SPI vs EMDA for 4 patients.

2.4 Deep Learning

In order to identify the optimal model architecture, an iterative approach is followed. Initially, random model architectures were tested to evaluate their performance. Based on the initial results, we made subsequent modifications to improve the model's performance. This study used data from 90 patients, 70 patients for training, 10 for testing, and 10 for validation. Detailed descriptions of the used LSTM models are listed in Table 1, and Fig. 5 provides visual representations of the architectures used.

Table 1 List of used deep learning models.

Model No.	Number of Layers/Hidden Units	LR/Epochs	Batch Size	Features
1	3LSTM-60s Timestep / 16-32-64 Neurons	0.001 / 20	64	SPI-PPGA-HBI
2	2LSTM-60s Timestep/ 16 - 32 Neurons	0.00001 / 15	256	SPI-PPGA-HBI
3	1LSTM-60s Timestep/8 Neurons	0.00001 / 30	32	SPI-PPGA-HBI

3 RESULTS

Since EMDA data are obtained from four medical doctors, we applied ICC test to each patient's assessments. The test results provided evidence that the doctors' ratings were consistent in most cases. Fig. 6 shows four examples from the test results. The confidence interval was narrow in all cases, suggesting that the true ICC value is precise. Fig. 7 shows the pain assessments for one patient. Wherein, ICC value of 0.54 is considered fair, however, looking at the

trends of the assessments, it is very clear that there is consistency, and they seem to have almost a similar trend.

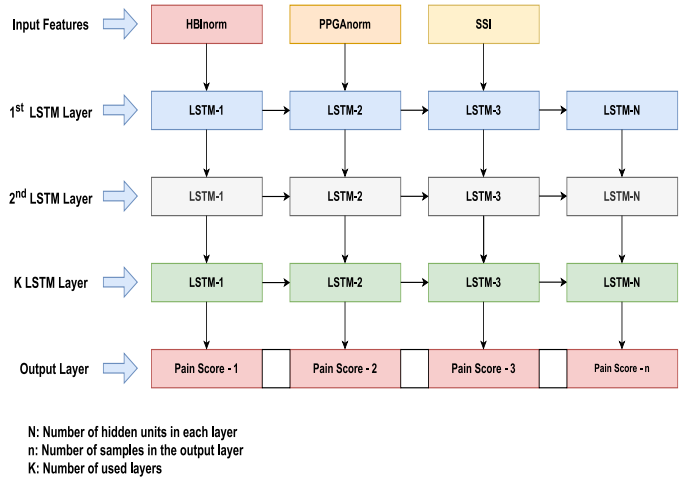

Fig. 5 Illustration of the LSTM architectures used in this study.

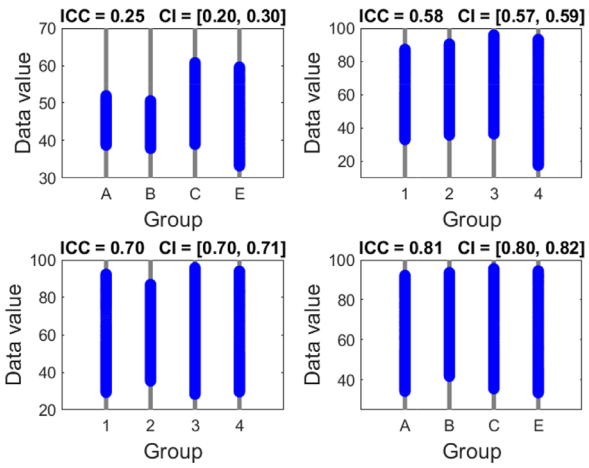

Fig. 6 ICC test results. CI represents the confidence interval.

Three models are developed to predict the pain scores using different hyperparameters. Performance metrics such as MAE and SD were used to assess the models' accuracy and variability, while also considering Corr to examine the relationship between the variables.

The first model consisted of three LSTM layers, and delivered poor test results, with an MAE of 6.9271±1.913, SD of 9.4635±2.456, and Corr of 0.5955±0.069. In contrast, the second model, with two LSTM layers, showed improved predictions, as demonstrated by an MAE of 3.418±0.715, SD of 3.847±0.557, and Corr of 0.6341±0.068.

The third model, which was built with one LSTM layer, was selected based on the stability of the training-validation loss curve, and its results were not significantly different from the 2-layer model, with an MAE of 3.4009±0.648, SD of 3.9099±0.548, and Corr of 0.6197±0.062. Fig. 8 illustrates the training and validation loss curves for all models.

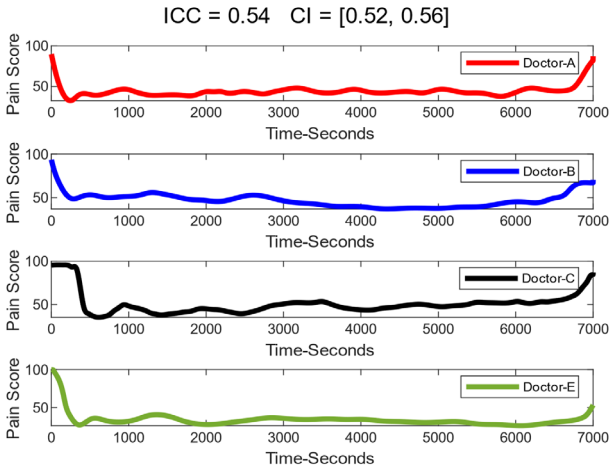

Fig. 7 Fair ICC value for one patient.

Once the models are trained, the optimal model is selected based on its statistical and performance metrics. Both the 2-layer and 3-layer models achieved comparable results. However, based on the evaluation metrics (MAE, SD, and Corr), the second model is determined to be the best. For cross-validation, a 5-fold model is applied. Since the data is formatted as time-series, shuffling the data is not possible. The model architecture for each fold is identical to that of the best model (i.e., the 2-layer model), consisting of two LSTM layers, 16 & 32 neurons, 60-second time steps, a learning rate of 0.00001, and a batch size of 256. Table 2 shows the statistical results for each patient.

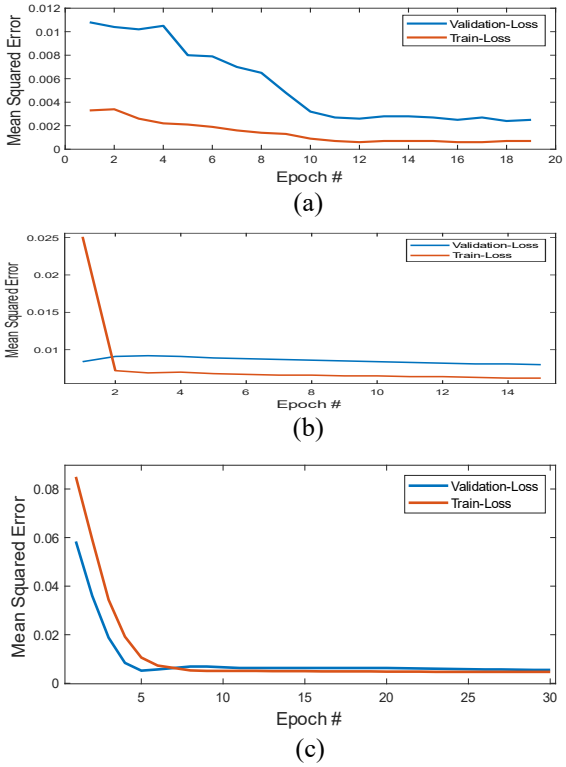

Fig. 8 Training-validation loss curves – (a) First model, (b) Second model, (c) Third model.

Table 2 Statistical results for cross-validation model (overall results for 5 folds).

Patient	SD	MAE	Corr
1	4.72732	3.8318	0.643
2	4.96698	3.36954	0.54168
3	3.22966	3.82178	0.63484
4	4.70286	4.23838	0.5229
5	5.77242	5.314	0.62778
6	5.57774	4.15002	0.62016
7	4.39652	3.15084	0.61552
8	4.74232	3.63476	0.58842
9	3.97634	4.55446	0.5137
10	5.13408	3.15744	0.66396
MEAN±SD	4.722±0.742	3.922±0.672	0.597±0.053

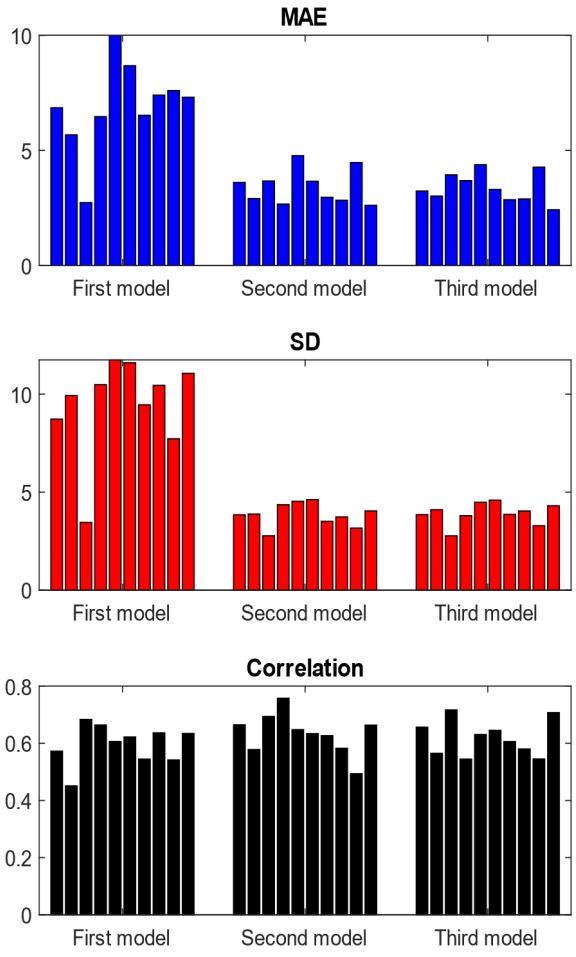

Fig. 9 Test statistical results.

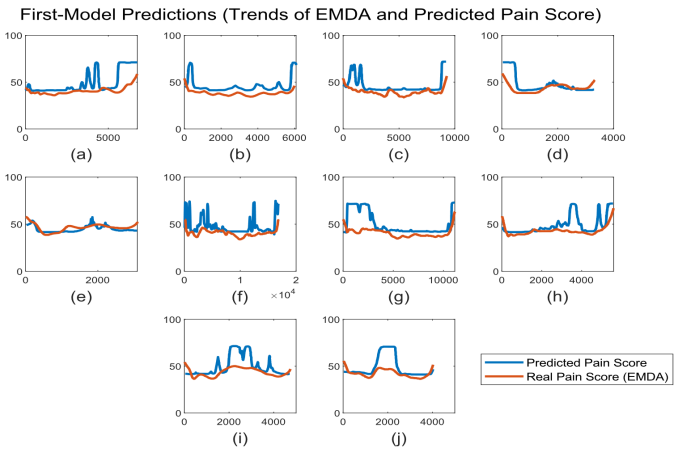

Fig. 10 First model predictions

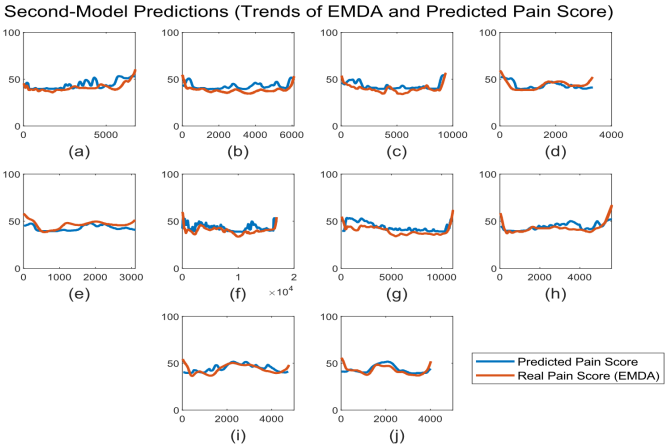

Fig. 11 Second model predictions

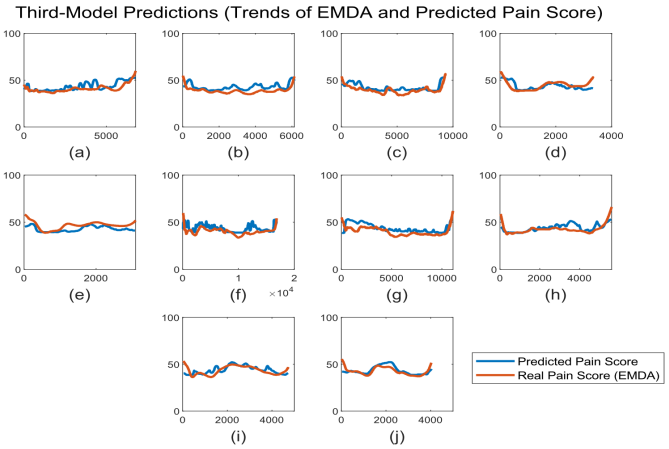

Fig. 12 Third model predictions

4 DISCUSSION AND CONCLUSION

SPI is evaluated based on PPGA and HBI, which relies on the activation of sympathetic and parasympathetic tones. Other factors are also known to affect those reactions, which in many cases leads to unreliable measures. Therefore, the aim of this study is to develop an LSTM

model to help predict pain scores during GA based on SPI and EMDA.

ICC test is used for its ability to measure the agreement or consistency between many raters. Unlike correlation, it considers the rater bias [20]. In this study, ICC test provided evidence that most of the patients' assessments has good to excellent agreement, however, in some cases where ICC result is poor or fair, the visual examination provided another fact, that even when the result is not good, the assessments are consistent, as shown in Fig. 7. Moreover, CI measures are narrow in most cases, which provides the true limits of ICC score, and a precise measure.

The data used in this study are collected in 2015. However, at that time, the hospital did not have the SSI algorithm. Therefore, the SPI is calculated based on the information provided in [3]. For many patients, SPI and EMDA have similar trends, despite the higher variability in SPI, which might not be related to pain, as shown in Fig. 4. High SPI indicates high pain, and low SPI indicates low pain, while SPI values falling between the upper and lower quarters (i.e., 25 and 75) indicate moderate pain.

The metrics used in this study are selected based on many factors. MAE provided the expected error value when testing new unseen data, which is low for the best model. SD is selected to test the variability within the prediction trends, as shown in Fig. 10, the first model had high variability, which should be lowerto follow the trend of EMDA. While the Corr metric provides the strength of agreement between the prediction and EMDA, All the developed models do not have a notable difference. Therefore, the best models are selected (i.e., The 2-layer model) based on MAE and SD.

Jean et al. [10] concluded that the relation between ANI and pain assessments is logically similar to this work. However, in their study, the trend of ANI is opposite to pain assessments. In comparison between this work findings and Jean et al. [10], their findings relied on one evaluator, which is MAE for both LSTM and MLP models. They achieved an MAE of 2.490 ± 0.522 for the best model (i.e., MLP model); on the other hand, while this work best model has achieved an MAE of 3.418 ± 0.7 15 (i.e., 2-layer LSTM model).

Additionally, this work models are evaluated based on MAE, SD, and Corr. Fig. 8 shows how unstable the first model was, where it has high validation loss, with inconsistent values, on the other hand, the second and third models were improved to have consistent and comparable loss values. Similarly, Fig. 10, Fig. 11, and Fig. 12 provide a clearer vision of test data, as the predictions are noisy and had high errors in the first model, in contrast, the second and third models have low error predictions.

As shown in Fig. 9, the Corr does not have a notable difference, while MAE and SD are remarkably improved for the second and third models. Although our MAE is slightly higher, it should be noted that in this work models, the balance between evaluator the results is very important. As we went through the training process, more than 3 models were trained, and lower MAE values are achieved; however, low MAE affected the regularization of the used models, resulting in flat prediction curves. Hence, these models are excluded to maintain the balance between all evaluators. Moreover, our predictions followed the trend of EMDA, even when the model is performing poorly (i.e., 3-layer model), which provides strong evidence that SPI could be used to predict pain score in GA.

A potential challenge to applying this model in real-time scenarios is the need for a large amount of data to accurately estimate the distribution parameters, which could be solved by using a large group of patients' data available as a baseline for normalization. Then, when new data are collected in real-time, a time delay of 1-5 minutes at the beginning of the surgery would provide HBI and PPGA time series every second, and thus could be used with the group distribution, then it will be normalized based on a prior knowledge distribution. Another possible solution is to use the SPI data provided by the GE Healthcare monitoring device if available, which also has a time delay depending on the SPI algorithm. Moreover, our findings show promising results to predict the pain score based on SPI, which might lead to another algorithm that does not need data normalization since the raw data used to calculate SPI are extracted from the PPG signals, which might be studied in the future.

In conclusion, SPI showed promising result in predicting pain scores during GA; however, future models improvements are necessary, including improving data consistency, and using different modelling approaches. In addition, the use of other clinical factors that influence the DoA, such as BP, skin conductance, or other biosignal features could improve the performance of the model.

ACKNOWLEDGMENTS

This research is supported by Ministry of Science and Technology (Grant number: MOST 110-2221-E-155 -004 -MY2).

REFERENCES

[1] C. Dodds, "General Anaesthesia," Drugs, vol. 58, no. 3, pp. 453–467, 1999, doi: 10.2165/00003495-199958030-00006.

[2] M. Gruenewald et al., "Influence of different remifentanil concentrations on the performance of the surgical stress index to detect a standardized painful stimulus during sevoflurane anaesthesia," British Journal of Anaesthesia, vol. 103, no. 4, pp. 586–593, Oct. 2009, doi: 10.1093/bja/aep206.

[3] M. Huiku et al., "Assessment of surgical stress during general anaesthesia," British Journal of Anaesthesia, vol. 98, no. 4, pp. 447–455, Apr. 2007, doi: 10.1093/bja/aem004.

[4] J. Wennervirta, M. Hynyneh, A.-M. Koivusalo, K. Uutela, M. Huiku, and A. Vakkuri, “Surgical stress index as a measure of nociception/antinociception balance during general anesthesia,” Acta Anaesthesiologica Scandinavica, vol. 52, no. 8, pp. 1038–1045, Aug. 2008, doi: 10.1111/j.1399-6576.2008.01687.x.

[5] X. Chen et al., “Comparison of Surgical Stress Index-guided Analgesia with Standard Clinical Practice during Routine General Anesthesia,” Anesthesiology, vol. 112, no. 5, pp. 1175–1183, May 2010, doi: 10.1097/aln.0b013e3181d3d641.

[6] M. M. R. F. Struys, C. Vanpeteghem, M. Huiku, K. Uutela, N. B. K. Blyaert, and E. P. Mortier, “Changes in a surgical stress index in response to standardized pain stimuli during propofol–remifentanil infusion,” British Journal of Anaesthesia, vol. 99, no. 3, pp. 359–367, Sep. 2007, doi: 10.1093/bja/aem173.

[7] J. Ahonen, R. Jokela, K. Uutela, and M. Huiku, “Surgical stress index reflects surgical stress in gynaecological laparoscopic day-case surgery,” British Journal of Anaesthesia, vol. 98, no. 4, pp. 456–461, Apr. 2007, doi: 10.1093/bja/aem035.

[8] Y. J. Won, B. G. Lim, Y. S. Kim, M. Lee, and H. Kim, “Usefulness of surgical pleth index-guided analgesia during general anesthesia: a systematic review and meta-analysis of randomized controlled trials,” Journal of International Medical Research, vol. 46, no. 11, pp. 4386–4398, Sep. 2018, doi: 10.1177/0300060518796749.

[9] M. Roy Chowdhury, R. Madanu, M. F. Abbod, S.-Z. Fan, and J.-S. Shieh, “Deep learning via ECG and PPG signals for prediction of depth of anesthesia,” Biomedical Signal Processing and Control, vol. 68, p. 102663, Jul. 2021, doi: 10.1016/j.bspc.2021.102663.

[10] W.-H. Jean, P. Sutikno, S.-Z. Fan, M. F. Abbod, and J.-S. Shieh, “Comparison of Deep Learning Algorithms in Predicting Expert Assessments of Pain Scores during Surgical Operations Using Analgesia Nociception Index,” Sensors, vol. 22, no. 15, p. 5496, Jul. 2022, doi: 10.3390/s22155496.

[11] M. Jeanne, R. Logier, J. De Jonckheere, and B. Tavernier, “Validation of a graphic measurement of heart rate variability to assess analgesia/nociception balance during general anesthesia,” Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2009, pp. 1840–1843, 2009, doi: 10.1109/IEMBS.2009.5332598.

[12] R. Logier, M. Jeanne, and B. Tavernier, “Method and device for assessing pain in human being,” University Hospital of Lille, vol. University of Lille II, 2004.

[13] Q. Liu et al., “Spectrum Analysis of EEG Signals Using CNN to Model Patient’s Consciousness Level Based on Anesthesiologists’ Experience,” IEEE Access, vol. 7, no. 99, pp. 53731–53742, Apr. 2019, doi: 10.1109/access.2019.2912273.

[14] Q. Liu, L. Ma, R.-C. Chiu, S.-Z. Fan, M. F. Abbod, and J.-S. Shieh, “HRV-derived data similarity and distribution index based on ensemble neural network for measuring depth of anaesthesia,” PeerJ, vol. 5, p. e4067, Nov. 2017, doi: 10.7717/peerj.4067.

[15] G. J. A. Jiang et al., “Sample Entropy Analysis of EEG Signals via Artificial Neural Networks to Model Patients’ Consciousness Level Based on Anesthesiologists Experience,” BioMed Research International, vol. 2015, pp. 1–8, 2015, doi: 10.1155/2015/343478.

[16] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, “An optimal filter for short photoplethysmogram signals,” Scientific Data, vol. 5, no. 1, May 2018, doi: 10.1038/sdata.2018.76.

[17] M. Huiku et al., “Assessment of surgical stress using heart rate and plethysmographic pulse wave amplitude variability,” cris.vtt.fi, 2005. https://cris.vtt.fi/en/publications/assessment-of-surgical-stress-using-heart-rate-and-plethysmograph

[18] V. Bonhomme et al., “Comparison of the Surgical Pleth Index TM with haemodynamic variables to assess nociception–anti-nociception balance during general anaesthesia,” British Journal of Anaesthesia, vol. 106, no. 1, pp. 101–111, Jan. 2011, doi: https://doi.org/10.1093/bja/aeq291.

[19] T. Ledowski, M. Schneider, M. Gruenewald, R.K. Goya, S.R. Teo, and J. Hruby, “Surgical pleth index: prospective validation of the score to predict moderate-to-severe postoperative pain”, Br J Anaesth. 2019 Aug; 123(2): e328–e332.

[20] J. Liu, W. Tang, G. Chen, Y. Lu, C. Feng, and X. M. Tu, “Correlation and agreement: overview and clarification of competing concepts and measures,” Shanghai Archives of Psychiatry, vol. 28, no. 2, pp. 115–120, Apr. 2016, doi: https://doi.org/10.11919/j.issn.1002-0829.216045.