Noname manuscript No.
(will be inserted by the editor)

Frame-wise Detection of Surgeon Stress Levels during Laparoscopic Training Using Kinematic Data

Yi Zheng[1] (ORCID:0000-0002-1830-0680) · Grey Leonard[2] · Herbert Zeh[2] · Ann Majewicz Fey[1,2] (ORCID:0000-0002-1802-6730)

Received: date / Accepted: date

Abstract Purpose Excessive stress experienced by the surgeon can have a negative effect on the surgeon's technical skills. The goal of this study is to evaluate and validate a deep learning framework for real-time detection of stressed surgical movements using kinematic data.

Methods 30 medical students were recruited as the subjects to perform a modified peg transfer task and were randomized into two groups, a control group (n=15) and a stressed group (n=15) that completed the task under deteriorating, simulated stressful conditions. To classify stressed movements, we first developed an attention-based Long-Short-Term-Memory recurrent neural network (LSTM) trained to classify normal/stressed trials and obtain the contribution of each data frame to the stress level classification. Next, we extracted the important frames from each trial and used another LSTM network to implement the frame-wise classification of normal and stressed movements.

Results The classification between normal and stressed trials using attention-based LSTM model reached an overall accuracy of 75.86% under Leave-One-User-Out (LOUO) cross-validation. The second LSTM classifier was able to distinguish between the typical normal and stressed movement with an accuracy of 74.96% with an 8-second observation under LOUO. Finally, the normal and stressed movements in stressed trials could be classified with the accuracy of 68.18% with a 16-second observation under LOUO.

Conclusion In this study, we extracted the movements which are more likely to be affected by stress and validated the feasibility of using LSTM and kinematic data for frame-wise detection of stress level during laparoscopic training. The proposed classifier could be potentially be integrated with robot-assisted surgery platforms for stress management purposes.

Keywords Laparoscopic surgery · Surgical stress · Deep learning · Motion analysis

Yi Zheng
E-mail: yi.zheng@austin.utexas.edu

[1] The Department of Mechanical Engineering, The University of Texas at Austin, TX 78712 USA
[2] The Department of Surgery, UT Southwestern Medical Center, TX 75390 USA

1 Introduction

Intra-operative surgical stress is commonly experienced by surgeons. Acute mental stress can compromise surgical skill and in turn, affect patient safety [3]. During laparoscopic procedures, it has also been shown that surgeons experience more stressful conditions than during open surgery due to limitations in visualization, work space volume, and an increased need for hand-eye coordination [5]. Performing laparoscopic surgery is a complex motor task. For complex tasks, it has been shown that external stressors can adversely affect motor performance [23]. The negative effects of external stress on surgical performance can include a higher number of errors made, less economy of motion, and increased completion time [2, 19, 12, 3].

Measuring Stress Level Excessive stress can have negative effects on a surgeon's technical skills, for example, leading to increased path length and a higher number of errors [19]. A common established method for measuring human stress levels involves the use of physiological data. Cortisol levels measured from saliva have been well studied as indicators of stress [3]. Heart rate, heart rate variability, and skin conductance level also can be used to quantify stress levels [6, 9, 25, 7]. However, these techniques can be time consuming, are relatively invasive, and may require surgeons to wear additional sensors on their bodies that may be cumbersome. Alternatively, in our previous studies, we validated the feasibility of using features extracted from kinematic data of the laparoscopic instrument tips (e.g., velocity, acceleration, and jerk) to distinguish between stressed and non-stressed conditions during laparoscopic training procedures using statistical analysis. These studies demonstrated that the kinematic data is a powerful tool for identifying stressed conditions. Additionally, kinematic data measuring techniques are less invasive than physiological sensing as they require fewer sensors that do not need to be worn by the surgeon [30, 15].

Demand for Real-time Detection of Stress Level Stress levels can vary during laparoscopic surgery and stress may come from different sources [1]. The aforementioned sensing techniques often measurements after the experimental trial. Continuous stress monitoring, however, could enable more granular stress-related data. For example, Weenk et al. [28] implemented continuous stress monitoring using a wearable sensor patch which monitored the heart rate variability (HRV) of surgeons. HRV analyis requires both time domain and frequency domain techniques, as well as collecting the baseline data from each subject, which can be computationally challenging. There is an important need to develop methods to detect stress levels in real-time during surgical procedures to help monitor surgeon performance and mitigate the potential risk to the patients.

More specifically, with the development of modern robotic-assisted surgical platforms, the kinematic data can be collected directly from robot joint encoders without additional sensors. The real-time detection of stress levels using kinematic data of surgical robot end-effectors can be integrated with the advanced control techniques on robotic-assisted surgical platforms to provide the surgeon with stress coping strategies.

Motivation for Recurrent Neural Networks Predictive modeling based on machine learning or deep learning methods has been widely used in the field of surgical skill assessment, such as k-Nearest Neighbors (kNN), logistic regression (LR) and support vector machines (SVM) [11, 26]. Wang et al. [27] used a convolutional neural network (CNN) architecture for real-time surgical skill assessment. These techniques used motion data as input and validated the fact that motion data can be used for characterizing surgical performance. For stress detection, Pandey [21] used several machine learning techniques (SVM, Logistic Regression) and heart rate as the input feature to predict patient acute stress condition.

With recent development in machine learning and deep learning, Recurrent Neural Networks (RNN), in particular, Long Short Term Memory (LSTM) models, have been shown to have important advantages in classifying and making predictions based on time-series data [13]. LSTM is an appropriate tool for temporal modeling and it is widely used in human activity recognition (HAR) and language processing due to its inherent structure to "memorize" and "forget" important points within a sequence of data [18, 20].

The advantages associated with handling time-series data using LSTM has attracted the attention of researchers in the field of surgical data science. DiPietro et al. [10] applied LSTM to joint segmentation and classification of surgical activities from robot kinematic data. Kannan et al. [14] presented a model of a combination of a convolutional neural network (CNN) and an LSTM network to process the video data for recognition of the type of a laparoscopic surgery (e.g. adrenalectomy, gastric bypass, cholecystectomy etc.).

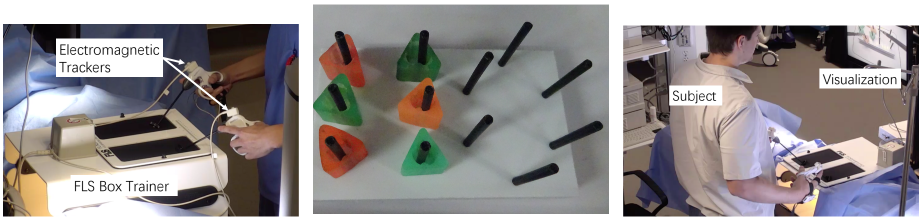
Recently, the attention mechanism has also been proposed for sequence modeling. Bahdanau et al. first introduced attention in machine translation where the output will focus its attention on a certain part of a sequence [4]. Neural networks have demonstrated performance improvements when integrated with an attention mechanism. Attention mechanisms has been widely used in variety of sequence modeling projects, such as machine translation [4, 29], sentiment classification [16], time-series prediction [22], etc.

As inspired by these studies, we decided to move a step forward to using predictive modeling techniques and kinematic data to implement a near real-time detection of surgical stress levels. Our hypothesis is that the surgeon's stress level during laparoscopic surgery can be extracted from the instrument handles movements within a short period of observation. In this study, we first implemented an attention-based LSTM classifier to classify normal/stressed trials as well as obtained the movements which were most affected by the stress. Then, we implemented another LSTM classifier to detect normal/stress movements based on the attention obtained from the first step.

2 Background and Preliminary Work

2.1 Experiment and Dataset

We used a portion of the dataset which came from one of our previous studies [15, 30]. 30 medical students (29 were right-handed and 1 was left-handed) at the University of Texas Southwestern Medical Center were recruited in this IRB approved study (UTD #14-57, UTSW STU #032015-053) and informed consent was obtained.



(a) Box Trainer and EM Trackers (b) FLS Peg Transfer Task (c) Subject and Visualization

Fig. 1: Simulator Setup.

After informed consent, each subject participated in a 10-minute tutorial on the Fundamentals of Laparoscopic Surgery (FLS) peg transfer drill to be familiarized with the instruments and the requirements of the experimental task. Subjects were randomly assigned into a control (n = 15) or stressed (n = 15) group.

During the experiment, each subject was asked to complete a 6-minute peg transfer task on a FLS box trainer in a high-fidelity simulated operating room (one trial per subject). The FLS box trainer was placed in the abdominal section of a medical manikin which was draped. A pair of electromagnetic (EM) trackers were used to capture the time-series data of motions (Fig. 1a). The EM trackers were mounted to the handles of the laparoscopic instruments. The data was recorded at a frequency of 256 Hz from the EM trackers.

The data collected by the EM trackers included x_h-, y_h-, z_h- positional coordinates in space and quaternions q_0-, q_1-, q_2-, q_3-. The position coordinates determined the instrument handle positions in space and the quaternions were used to determine the rotation matrix for calculating the 3 dimensional instrument tip positions (x_t-, y_t-, z_t-). The instrument tip positions were calculated using handle positions, a rigid body transformation obtained by quarternions and the instrument geometry. Both instrument handle and instrument tip positions were saved in the dataset.

The stressors in the study included the vital signs of the medical manikin and the moderator's feedback during the task. In control group, each subject proceeded while hearing normal vital signs and with no feedback from the moderator. What is worthy mentioning is, in stressed group, each subject performed the task under a period of progressively deteriorating vital signs, with a particular increase in intensity beginning at the 3-minute mark. The moderator also provided feedback to the stressed subject and the feedback culminated in 30 seconds of cardiac arrest and the expiration of the medical manikin.

Besides the kinematic data from EM trackers, other data was collected and evaluated through video review, such as number of pegs transferred, number of errors made. Additionally, a blinded, independent reviewer with training in OSATS scoring graded each subject using a modified OSATS (mOSATS) rubric [17]. The subjects were also brought to complete the State-Trait-Anxiety-Inventory (STAI) to measure subjective stress after the experiment [24].

Overall, in this study, we only used the kinematic portion of our previously collected dataset. The dataset in this study contains the time-series 3-D positional data of both instrument handles (x_h-, y_h-, z_h-) of each subject throughout the 6-minute peg transfer task. We removed the data of one subject (in control group, right-handed) due to sensor failure during experiment. We down-sampled the data to 5Hz and organized the data of both instrument handles based on each subject's handedness, so the overall dataset of 29 subjects resulted in approximately 52,200 samples of six features x_{ND}, y_{ND}, z_{ND}, x_D, y_D, z_D (the subscript D is Dominant hand and ND is Nondominant hand).

2.2 Previous Results

In our previous studies, we calculated the kinematic metrics of the instrument tips, such as velocity, acceleration, and jerk. We also analyzed the scores obtained by mOSATS and STAI. Statistical analysis comparing the metrics between control and stressed groups was conducted.

According to our previous studies evaluating the experimental data, in general, the stressed group had higher velocity, acceleration, jerk, indicating less smooth movements on instrument tips; Smaller numbers of pegs transferred, larger numbers of error made, lower mOSATS scores and higher scores for the change from baseline (trait) to during the scenario (state) in STAI.

The significant differences between control and stressed groups in our previous studies indicated that kinematic data can be related to increased stress levels. The detailed results of these evaluations can be found in our previous studies [15, 30].

3 Methods

3.1 Trial-wise Classification and Attention

It was not known if all movements made by the subject within a trial would have been affected by the external stress. The goal of this step is to find the importance of each time step within a trial that contributes to the stress representation. In other words, we want to extract the movements that are more significantly affected by the stress.

The architecture of the proposed attention-based LSTM classifier is shown in Fig. 2. The input sequence $\{x_1, x_2, ..., x_T\}$ was the kinematic data of each trial. As mentioned in Section 3.1, the input kinematic data contains six features of the 3D positional data of both instrument handles (x_{ND}, y_{ND}, z_{ND}, x_D, y_D, z_D). For each input:

$$x_i = [x_{NDi}, y_{NDi}, z_{NDi}, x_{Di}, y_{Di}, z_{Di}]^T, i = 1...T \qquad (1)$$

The subscript D is dominant hand side and ND is dondominant hand side. The ground-truth label $y = \{0 \text{ or } 1\}$ was assigned to be control (normal) or stressed trials. The input sequence $\{x_1, x_2, ..., x_T\}$ was then fed into a Bidirectional LSTM to get the hidden state sequence $\mathbf{h} = \{h_1, h_2, ..., h_T\}$.

Then we measured the importance of each time step by computing a *tanh* function of hidden states $\mathbf{h}$:

$$\mathbf{e} = tanh(\mathbf{h}) = tanh(h_1, h_2, ..., h_T) \qquad (2)$$

e is called “energy” which can be interpreted as the contribution of the time step to the final representation of stress levels. The attention weights α_i were obtained by passing e_i to a $Softmax$ function, where ensured all attention weights of a trial sum to 1.

$$\alpha_i = \frac{exp(e_i)}{\sum_{i=1}^{n} exp(e_i)} \tag{3}$$

The attention weight α_i indicates how much attention the ground-truth label y should pay to the i^{th} time step. Then we can calculate the context vector as a weighted linear combination of all hidden states **h**:

$$context = \sum_{i=1}^{n} \alpha_i h_i \tag{4}$$

Finally, two fully connected layers with activation functions of $ReLU$ and $Softmax$ were added. The context vector passed through the final layers and gave a prediction of $\hat{y}$.

Through model training and testing, we obtained the attention vector of each trial which was able to tell us which time steps were more important for classifying the trials as control (normal) or stressed.

3.2 Movement Extraction

After obtaining the attention vector of each trial, we used a sliding window with a 50% overlap to organize the attention sequence and the input sequence into frames (Fig. 3). For each trial, we calculated the sum of each attention frame:

$$A_t = \sum(\alpha_i, \alpha_{i+1}, ..., \alpha_{i+m-1}) \tag{5}$$

m is the frame length. We also tested the performance of frame-wise classifier with different frame lengths (1s, 2s, 4s, etc.) in the following sections.

Then we calculated the third quartile of all A_t's in a trial as the threshold:

$$threshold = Q_3(A_1, A_2, ..., A_n) \tag{6}$$

n is the number of frames for each trial. Q_3 is the third quartile. We considered any frame with an $A_t > threshold$ to be “important” to reflect the effect of stress.

More specifically, a frame with an $A_t > threshold$ in a control (normal) trial was considered to be a “representative” normal movement. Similarly, a frame with an $A_t > threshold$ in a stressed trial was considered to be a “representative” stressed movement.

Then, a subset of the original dataset containing the “representative” normal and stressed movements could be extracted based on the “important” frames for further classification (Fig. 2).

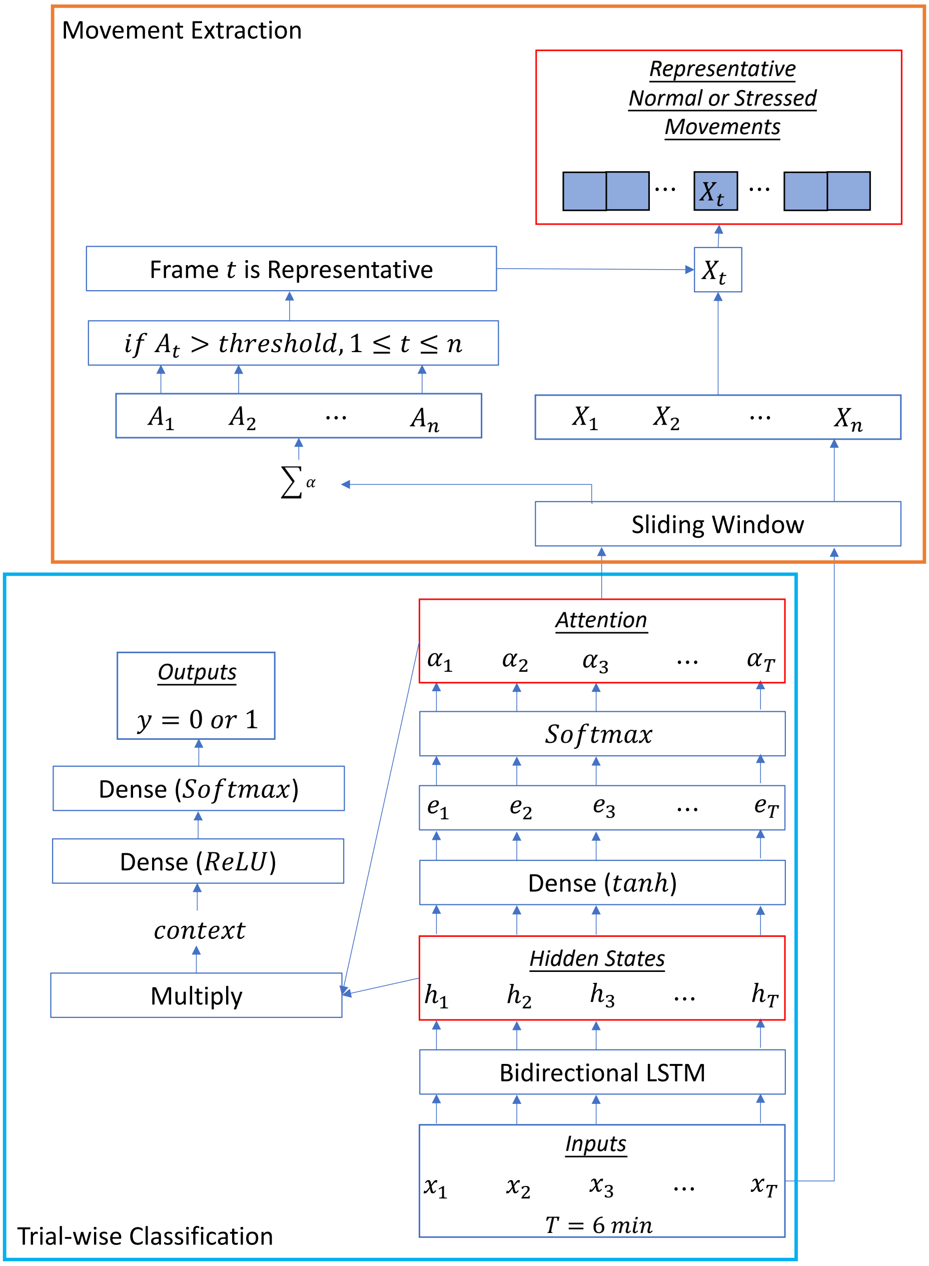

Movement Extraction
Representative Normal or Stressed Movements
X_t
Frame t is Representative
X_t
$if\ A_t > threshold, 1 \leq t \leq n$
A_1 A_2 ⋯ A_n
X_1 X_2 ⋯ X_n
$\sum \alpha$
Sliding Window
Attention
α_1 α_2 α_3 ⋯ α_T
Outputs
$y = 0\ or\ 1$
$Softmax$
Dense ($Softmax$)
e_1 e_2 e_3 ⋯ e_T
Dense ($ReLU$)
Dense ($tanh$)
$context$
Hidden States
Multiply
h_1 h_2 h_3 ⋯ h_T
Bidirectional LSTM
Inputs
x_1 x_2 x_3 ⋯ x_T
$T = 6\ min$
Trial-wise Classification

Fig. 2: Model architecture of attention-based LSTM classifier for trial-wise classification and movement extraction based on attention.

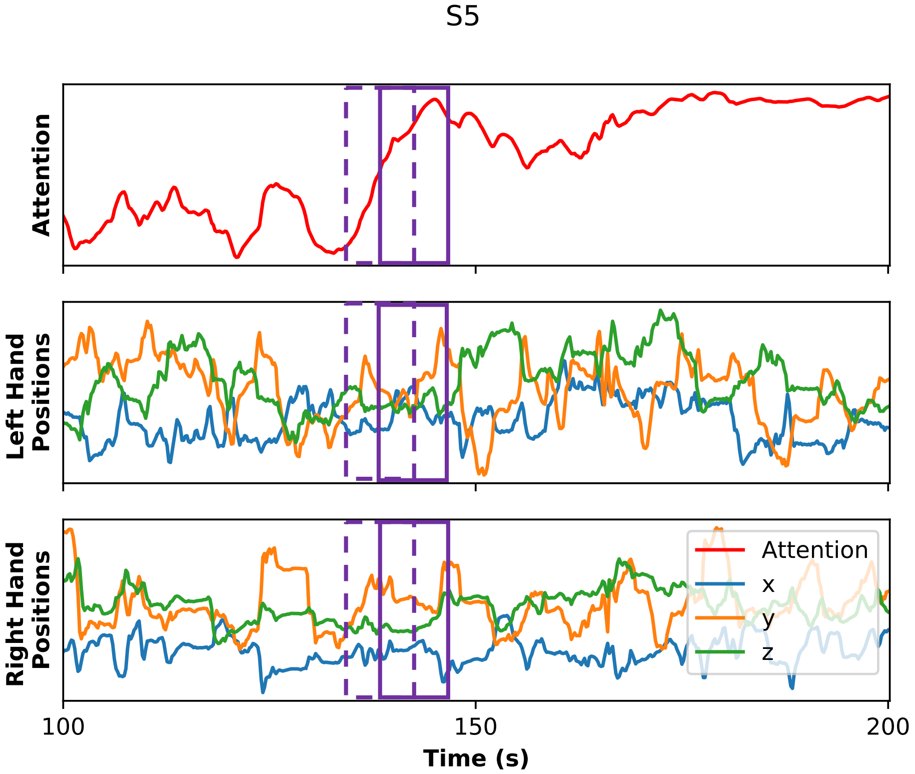


Fig. 3: Example of using sliding windows to organize the sequential data. The purple rectangles indicate the frames and an overlap of 50% between dashed (frame $t-1$) and solid (frame t) rectangles.

3.3 Frame-wise Classification

The training dataset of frame-wise classification is the "representative" normal and stressed movements extracted from Section 3.2.

The frame-wise classifier is a simple LSTM classifier which has an LSTM layer, a fully connected layer with the activation function of *ReLU* and a fully connected layer with the activation function of $softmax$ to output the probability of a given data frame belonging to each of the 2 stress levels (normal or stressed).

We implemented the architectures of both models using Keras library based on Python 3.7 [8]. We tested the hyperparameters of the proposed networks by trial-and-error. The models were trained by minimizing the categorical cross entropy loss function between the predicted and ground-truth labels at a *learning rate* of 0.001, first and second momentum of 0.9 and 0.999, and weight decay of 10^{-8}.

3.4 Model Training and Validation

It is a standard practice to test the model by leaving aside a portion of the data as testing dataset, using the remaining portion for training. To evaluate the performance of our proposed classifiers, we adopted Leave-One-User-Out cross-validation (LOUO). We used LOUO to test if the classifiers were generalized enough for unseen data. Our LOUO used the i^{th} subject as testing dataset and the rest for training, and iterated throughout all the 29 subjects. The mean values of all 29 iterations' performance metrics were reported and will be shown in the following sections.

3.5 Performance Metrics

In classification, there are four common metrics for evaluating the performance of a classifier - Accuracy, Precision, Recall and F1-score. Accuracy is the ratio of correct predictions $(T_p + T_n)$ to the total predictions $(T_p + F_p + T_n + F_n)$; Precision is the ratio of correct positive predictions (T_p) to the total positive results $(T_p + F_p)$ predicted by the classifier; Recall is the ratio of correct positive predictions (T_p) to the total actual results $(T_p + F_n)$. F1-score is a measure of a classifier's accuracy which takes the harmonic mean of the precision and recall.

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + T_n + F_n}, \tag{7}$$

$$Precision = \frac{T_p}{T_p + F_p}, \tag{8}$$

$$Recall = \frac{T_p}{T_p + F_n}, \tag{9}$$

$$F1 - score = \frac{2(Recall * Precision)}{Recall + Precision}. \tag{10}$$

4 Results

To test the effectiveness of the proposed methods, we conducted the following analysis: (1) we evaluated the performance of our attention-based trial-wise classifier for evaluating the stress level of each trial; (2) we validated the attention vectors that were obtained from trial-wise classification and interpreted the practical meaning of attention based on the experimental designs; (3) we extracted the "representative" movements based on the attention vectors, and tested if these extracted movements were able to train the frame-wise classifier for detecting normal and stressed movements.

4.1 Trial-wise Classification and Attention

According to the experiment, each subject finished one 6-minute peg transfer trial under either control (normal) condition or stressed condition. We remove the data

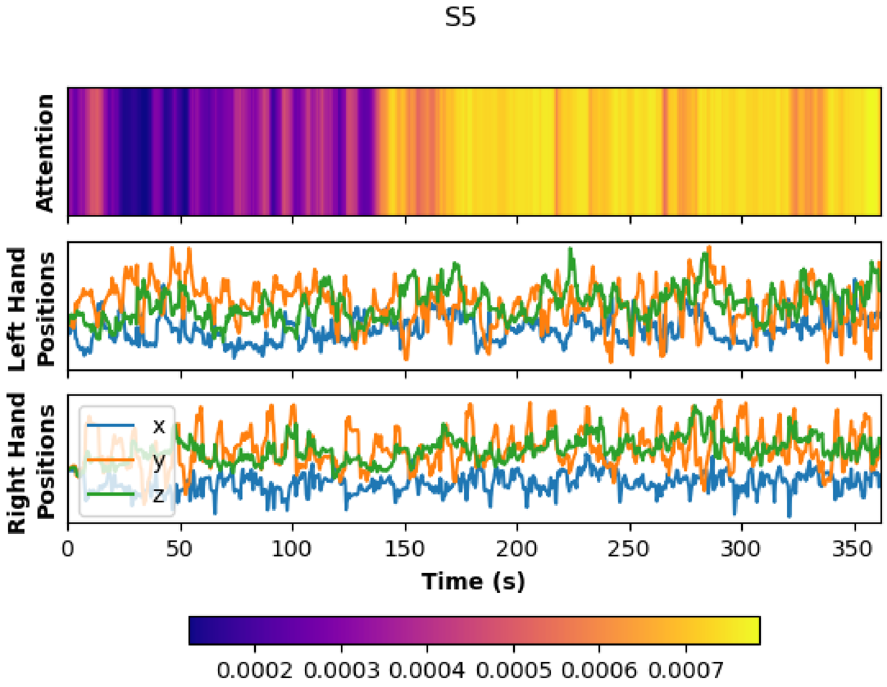


Fig. 4: Visualization of attention of an example stressed trial. Top: a heat map colorizes the magnitude of attention at each time step. Bottom: the time-series positions of both instrument handles.

of one subject from the control group (right-handed) due to sensor failure during experiment, therefore resulting in a dataset of 14 subjects (or trials) in control group and 15 subjects (or trials) in stressed group.

First, we implemented the attention-based LSTM classifier to distinguish between control (normal) and stressed trials. We annotated the control (normal) trials as "0" and stressed trials as "1". The input data was the kinematic data of each trial. After hyperparameter tuning, we obtained the performance metrics of this classifier under LOUO cross-validation scheme (Accuracy: 75.86%, Precision: 75.48%, Recall: 77.02%, F1-score: 76.24%).

In addition, we also obtained the attention vector of each trial which indicated the contribution of each time step to the classification. We used the sliding-window to organized the attention into frames. The sum of attention of each frame was computed. The frames which had an attention sum greater than the 3rd quartile in each trial were considered to be representative normal or stressed movements.

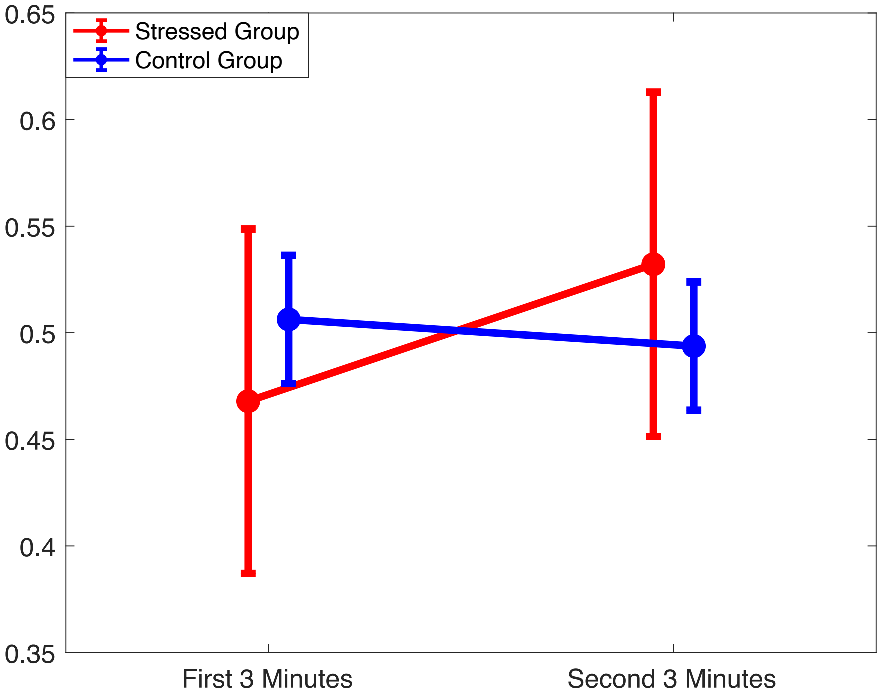


Fig. 5: Comparison between the attention of first and second 3 minutes in control (normal) and stressed trials. The second 3 minutes in stressed trials are associated with higher attention.

4.2 Validation of Attention Mechanism

We also divided the attention vector in stressed trials into first-3-minute and second-3-minute halves. We took the attention sums of these two halves and ran the ANOVA test. The results showed that the attention sum of the second half in stressed trials was significantly greater than the attention sum of the first half in stressed trials ($p = 0.0386$), which means the movements in second half contributed more to the classification of "stressed" and were more affected by the stressors.

The same experiment was also conducted on the attention vector in control trials. The results showed that the attention sums of the first and second 3 minutes in control trials were not significantly different ($p = 0.2812$), as shown in Fig. 5.

This finding is also consistent with our experimental design: the stressed group experienced increasingly intensive stressors in the second 3 minutes of each trial, therefore, validating the feasibility of the attention mechanism in this study.

Table 1: Performance summary of classification between "representative" normal and stressed movements under different frame sizes using LOUO cross-validation. Bold column denotes the best results.

Metrics	1s	2s	4s	**8s**	16s
Accuracy	60.91	64.73	72.24	**74.96**	70.85
Precision	60.92	64.70	72.21	**75.03**	71.21
Recall	60.93	64.59	72.22	**75.04**	71.04
F1-score	60.93	64.65	72.22	**75.04**	71.13

4.3 Movement Extraction and Classification

We implemented another simple LSTM model to classify the representative normal and stressed movements extracted from each trial based on attention. The training dataset contained the representative (high-attention) frames in control and stressed groups. Any frame had an attention sum greater than the 3rd quartile in a control trial was considered to be representative normal movements and any frame had an attention sum greater than the 3rd quartile in a stressed trial was considered to be representative stressed movements.

The frame sizes in classification using data streams play an important role as they need to contain enough information. In order to optimize the performance of our classifier, we repeated the training and LOUO cross-validation process with the data of four different frame sizes (1s, 2s, 4s, 8s, 16s).

Under LOUO cross-validation, the classification performance metrics were obtained. The frame size of 8 seconds showed the best results, as shown in Table 1: (Accuracy: 74.96%, Precision: 75.03%, Recall: 75.04%, F1-score: 75.04%).

4.4 Frame-wise Classification in Stressed Trials

As we mentioned in previous sections, the movements are not equally affected by the stressful condition which means that normal movements can still exist while the surgeon operating under stress. We have extracted "representative" normal and stressed movements from both control and stressed groups based on the attention vectors, and validated a classifier that could be used to distinguish between normal and stressed movements in Section 4.3. For this step, we test if these "representative" movements are applicable to classification between different movements in stressed trials.

The training dataset contains the normal and stressed movements extracted from control and stressed trials, as mentioned in Section 4.3.

The testing dataset only contained the data of stressed trials. For ground-truth labeling in stressed trial (Fig 6), we annotated the frame which had an attention sum greater than the third quartile of all attention sums in a trial as "stressed (1)", and the frame which had an attention sum less than the first quartile of all attention sums in a trial as "normal (0)".

We used the LOUO cross-validation to test the performance of frame-wise classifier. The i^th subject in stressed trial was used for testing. The training dataset should not include the data of the i^th subject. And the same process iterated throughout all 15 subjects in stressed group.

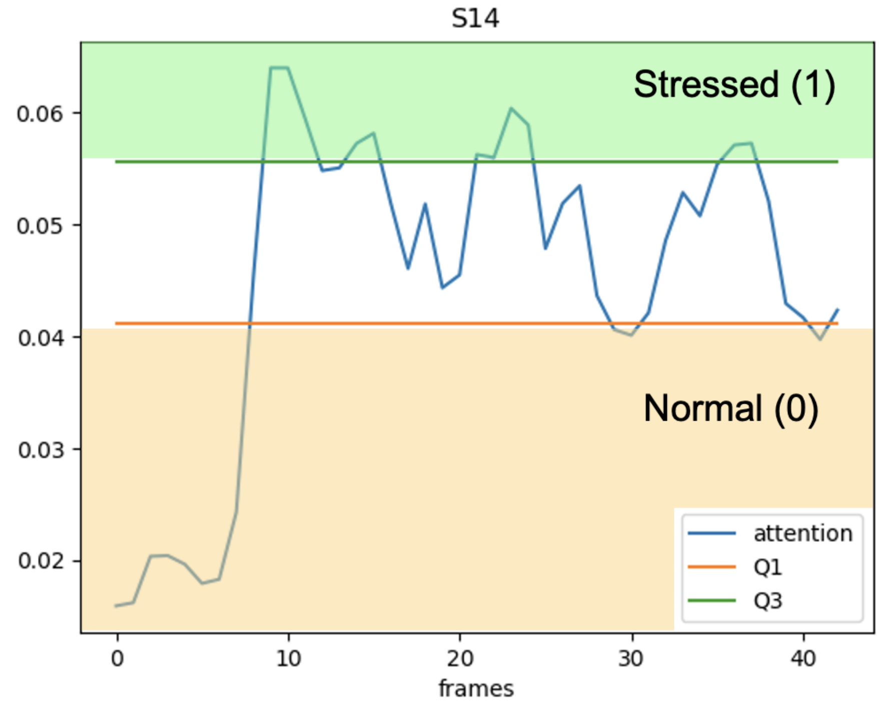


Fig. 6: Example of ground-truth labeling for stressed trials with a frame size of 8 second and an overlap of 50% (subject 14).

Table 2: Performance summary of classification between normal and stressed movements in stressed trials under different frame sizes using LOUO cross-validation. Bold column denotes the best results.

Metrics	1s	2s	4s	8s	**16s**
Accuracy	61.46	65.33	65.08	66.77	**68.18**
Precision	61.51	65.33	65.26	67.01	**68.30**
Recall	61.46	65.33	65.09	66.77	**68.18**
F1-score	61.48	65.33	65.17	66.89	**68.24**

The LOUO cross-validation results are summarized in Table 2. The frame size of 16 seconds showed the best results (Accuracy: 68.18%, Precision: 68.30%, Recall: 68.18%, F1-score: 68.24%).

5 Discussion

Although many studies have been investigated surgeon stress levels and cognitive load during training, none of these studies have implemented stress detection in near real-time, to our knowledge. Prior studies have also included the recording and analysis of physiological data, for example, heart rate, heart rate variability, eye movements and skin conductance level in ways that can reflect subject stress levels directly; however, these methods require external sensors and are also not real-time. The goal of our study is to validate the feasibility of a neural network approach to enable near-real time stress level detection using only kinematic data.

LSTM Recurrent Neural Networks have been widely used for prediction with time-series data as the input. More specifically, the LSTM with attention mechanism has gained its popularity recently in the field of sequence to sequence (seq2seq) modeling, such as machine translation and semantic analysis. We started with an attention-based LSTM architecture to distinguish between the control (normal) and stressed trials as well as getting the attention vector for movement extraction and used another simple LSTM classifier to distinguish normal and stressed movements.

We validated our classifiers using a common cross-validation method: LOUO cross-validation. The goal of LOUO cross-validation is to test if the model is generalized for unseen data, i.e. having a high accuracy with the data from a new (unseen) subject. For trial-wise classification, we obtained the accuracy of 75.86% under LOUO as well as the attention vector of each trial.

In terms of the frame sizes, we tested different frame sizes (1s, 2s, 4s, 8s, 16s) for frame-wise classification. A larger frame size can have an improved performance in classification. But the classifier performance decreases when the frame size continuously increases due to the fact that the LSTM can face challenges when handling longer sequences. Our proposed frame-wise classifier was able to distinguish between the "representative" normal and stressed movements with an accuracy of 74.96%; and an accuracy of 68.18% when the frame-wise classifier was applied to detecting normal and stressed movements within the stressed trials.

One limitation of this study is that we only tested a fixed size data frame. However, a surgical procedure consists of different surgical gestures, for example, moving, lifting and grasping, with different lengths of time period. Different kinds of surgical gestures could be affected by the surgical stressors differently. One direction of our future work is to overlap the attention vector on the recorded video, and extracted the surgical gestures that are significantly affected by the stressors. The second limitation of this study is the number of features. We only had 3D positional data as the input (x_{ND}, y_{ND}, z_{ND}, x_D, y_D, z_D). Especially, when we transplant this method to robot-assisted surgical systems where more information can be streamed, for example, rotation matrix, linear velocities and angular velocities, recruiting a variety of kinematic data may help improve the overall performance of our proposed method. Another limitation of the experiment is the lack of expertise levels and baseline data collection. We only had medical students recruited and only one trial (control or stressed) for each subject in the study. A better generalization can be made if subjects included attending, fellow, and resident surgeons in a large number, as well as baseline trials prior to the experiment to wash out the individual's inherent psychomotor skills.

It is worthy noting that we used the kinematic data on instrument handles (x_h-, y_h-, z_h-) in this study. There are several reasons why we used the data on instrument handles: First, handles motion could better capture the hands motion as shown in Fig. 1a; Second, our long term goal is to provide stress coping strategies on robot-assisted surgical platforms where we can provide haptics on surgeon-side manipulators based on the kinematic data of hands motion. Therefore, one direction of future work is to conduct a similar experiment using a robot-assisted surgical platform, such as da Vinci Research Kit (dVRK), to study the differences of identifying stressed conditions between conventional laparoscopic surgery and robot-assisted laparoscopic surgery.

6 Conclusion

In this study, we developed a deep learning model to extract and detect stressed movements from kinematic data during laparoscopic surgical training tasks. We first validated an attention-based LSTM model for classification of normal/stressed surgical training trials. Based on the attention, we were able to extract the typical movements that contributed to the classification of each trial. Finally, we validated another simple LSTM classifier and we were able to distinguish between the normal and stressed movements using a short period of data. We tested the model under LOUO cross-validation scheme, and it showed that the model was generalized to unseen data.

Our proposed method has the following advantages for surgical stress detection: First, only kinematic data was used. Unlike physiological sensing techniques, kinematic sensing does not require the subject to wear sensors, especially in robot-assisted surgical systems. Second, our frame-wise classifier takes a short period of movement as input and outputs its stress level. This frame-wise classification enables near real-time detection of stress level during surgical procedures. Finally, our model avoids feature extraction prior to feeding data to the model. Using the raw data can potentially expedite detection to near real-time.

Our proposed model has the ability of high accuracy and fast computational speed which is suitable for near real-time detection of surgical stress level using kinematic data. Future experiments should be done to study the detection of stress on a robot-assisted surgical platform due to the inherent differences between convention laparoscopic surgery and robot-assisted laparoscopic surgery, for example, motion scaling and fulcrum effects. We believe that this study paved way for continued research on mitigating the negative effect of surgical stress on robot-assisted surgical systems where the kinematic data can be streamed directly.

Acknowledgements The authors would like to thank the UTSW Training Resident Doctors As Innovators in Science (TARDIS) program, supporting co-Author Grey Leonard. We would like to thank the UTSW Simulation Center director and staff for their invaluable assistance with data collection. We also thank Dr. Chang Su at Weill Cornell Medical College, Cornell University, for his help with understanding deep learning architectures.

7 Declarations

Funding This work was partially supported by the National Science Foundation (NSF) grant number #1846726 and the National Institutes of Health (NIH) grant number #1R01EB030125-01. The UTSW Training Resident Doctors As Innovators in Science (TARDIS) program financially supported co-Author Grey Leonard.

Conflicts of Interest The authors declared that they have no conflict of interest.

Availability of Data and Material Data used in this study could be made available by request.

Code Availability The analysis code used in this study could be made available by request.

Ethics Approval This experiment was conducted using ethical practices in accordance with the University of Texas at Dallas (#14-57) and UTSW IRB offices (STU #032015-053).

Consent to Participate Written consent was obtained from the subjects to participate in this study, following IRB guidelines.

Consent to Publish The consent form for this study included language indicating that this research would potentially lead to a scientific publication.

References

1. Anton NE, Montero PN, Howley LD, Brown C, Stefanidis D (2015) What stress coping strategies are surgeons relying upon during surgery? In: American Journal of Surgery, Elsevier Inc., vol 210, pp 846–851
2. Arora S, Sevdalis N, Nestel D, Tierney T, Woloshynowych M, Kneebone R (2009) Managing intraoperative stress: what do surgeons want from a crisis training program? American Journal of Surgery 197(4):537–543, DOI 10.1016/j.amjsurg.2008.02.009
3. Arora S, Sevdalis N, Aggarwal R, Sirimanna P, Darzi A, Kneebone R (2010) Stress impairs psychomotor performance in novice laparoscopic surgeons. Surgical Endoscopy 24(10):2588–2593
4. Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 1409.0473
5. Berguer R, Smith WD, Chung YH (2001) Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. Surgical Endoscopy 15(10):1204–1207
6. Böhm B, Rötting N, Schwenk W, Grebe S, Mansmann U (2001) A prospective randomized trial on heart rate variability of the surgical team during laparoscopic and conventional sigmoid resection. Archives of Surgery 136(3):305–310
7. Boucsein W (2012) Electrodermal activity: Second edition, vol 9781461411. Springer US
8. Chollet F (2015) Keras. https://github.com/fchollet/keras
9. Czyzewska E, Kiczka K, Czarnecki A, Pokinko P (1983) The surgeon's mental load during decision making at various stages of operations. European Journal of Applied Physiology and Occupational Physiology 51(3):441–446
10. DiPietro R, Lea C, Malpani A, Ahmidi N, Vedula SS, Lee GI, Lee MR, Hager GD (2016) Recognizing Surgical Activities with Recurrent Neural Networks. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, pp 551–558
11. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD (2018) Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery 14(1):e1850, URL http://doi.wiley.com/10.1002/rcs.1850

12. Goodell KH, Cao CG, Schwaitzberg SD (2006) Effects of cognitive distraction on performance of laparoscopic surgical tasks. Journal of Laparoendoscopic and Advanced Surgical Techniques 16(2):94–98, DOI 10.1089/lap.2006.16.94
13. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9(8):1735–1780
14. Kannan S, Yengera G, Mutter D, Marescaux J, Padoy N (2020) Future-State Predicting LSTM for Early Surgery Type Recognition. IEEE Transactions on Medical Imaging 39(3):556–566, DOI 10.1109/TMI.2019.2931158, 1811.11727
15. Leonard G, Cao J, Scielzo S, Zheng Y, Tellez J, Zeh HJ, Fey AM (2020) The Effect of Stress and Conscientiousness on Simulated Surgical Performance in Unbalanced Groups: A Bayesian Hierarchical Model. Journal of the American College of Surgeons 231(4):S258, DOI 10.1016/j.jamcollsurg.2020.07.397
16. Ma D, Li S, Zhang X, Wang H (2017) Interactive Attention Networks for Aspect-Level Sentiment Classification. Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018 pp 25–30, URL http://arxiv.org/abs/1709.00893, 1709.00893
17. Martin JA, Regehr G, Reznich R, Macrae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery 84(2):273–278, DOI 10.1046/j.1365-2168.1997.02502.x
18. Milenkoski M, Trivodaliev K, Kalajdziski S, Jovanov M, Stojkoska BR (2018) Real time human activity recognition on smartphones using LSTM networks. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings, Institute of Electrical and Electronics Engineers Inc., pp 1126–1131
19. Moorthy K, Munz Y, Dosis A, Bann S, Darzi A (2003) The effect of stress-inducing conditions on the performance of a laparoscopic task. Surgical Endoscopy and Other Interventional Techniques 17(9):1481–1484
20. Nammous MK, Saeed K (2019) Natural language processing: Speaker, language, and gender identification with LSTM. In: Advances in Intelligent Systems and Computing, Springer Verlag, vol 883, pp 143–156, URL https://doi.org/10.1007/978-981-13-3702-4_9
21. Pandey PS (2017) Machine Learning and IoT for prediction and detection of stress. In: Proceedings of the 2017 17th International Conference on Computational Science and Its Applications, ICCSA 2017, Institute of Electrical and Electronics Engineers Inc., DOI 10.1109/ICCSA.2017.8000018
22. Qin Y, Feyzabadi S, Allan M, Burdick JW, Azizian M (2020) daVinciNet: Joint Prediction of Motion and Surgical State in Robot-Assisted Surgery. arXiv URL http://arxiv.org/abs/2009.11937, 2009.11937
23. Ryan ED (1962) Effects of stress on motor performance and learning. Research Quarterly American Association for Health, Physical Education and Recreation 33(1):111–119
24. Sielberger C, Gorsuch R, Vagg P, Jacobs G (1983) Manual for the state-trait anxiety inventory (form y)
25. Tendulkar AP, Victorino GP, Chong TJ, Bullard MK, Liu TH, Harken AH (2005) Quantification of surgical resident stress "on call". Journal of the American College of Surgeons 201(4):560–564
26. Vedula SS, Malpani A, Ahmidi N, Khudanpur S, Hager G, Chen CCG (2016) Task-Level vs. Segment-Level Quantitative Metrics for Surgical Skill Assess-

ment. Journal of Surgical Education 73(3):482–489

27. Wang Z, Majewicz Fey A (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International Journal of Computer Assisted Radiology and Surgery 13(12):1959–1970, URL https://doi.org/10.1007/s11548-018-1860-1, 1806.05796
28. Weenk M, Alken AP, Engelen LJ, Bredie SJ, van de Belt TH, van Goor H (2018) Stress measurement in surgeons and residents using a smart patch. American Journal of Surgery 216(2):361–368
29. Zhang B, Xiong D, Su J (2020) Neural Machine Translation with Deep Attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(1):154–163, DOI 10.1109/TPAMI.2018.2876404
30. Zheng Y, Leonard G, Tellez J, Zeh H, Fey AM (2021) Identifying kinematic markers associated with intraoperative stress during surgical training tasks. In: 2021 International Symposium on Medical Robotics (ISMR), IEEE, pp 1–7