**Results**

**Force Profile Segmentation**

Chart

Description automatically generatedA picture containing chart

Description automatically generated

1. (b)

**Fig. S1 |** **Accuracy and loss function values during training and validation steps for force profile segmentation using U-Net model.** History results for model accuracy and loss function value over 50 epochs overlaid for both training and validation iteration. a) The minimum validation loss function value occurred at epoch 28 and was 0.0878 (training loss = 0.0827). b) The historical accuracy for training has a consistent improvement over the trials and achieved 0.98 in training and 0.97 in validation.

Chart, treemap chart

Description automatically generated

**Fig. S2 | The summary prediction results for the segmentation model as binary classification problem using U-Net model.** Across the approximately 398K records of force data points with 54.4% in *ON* labels and 45.6% in *OFF* labels after the implementation of our rule-based data balancing,the accuracy of classification was 0.95, sensitivity (True Positive Rate: TPR) was 0.96, and specificity (True Negative Rate: TNR) was 0.94.

Chart, line chart

Description automatically generated

(a)

Chart, line chart

Description automatically generated

(b)

**Fig. S3 | The prediction performance for the segmentation model through receiver operating characteristic curves.** The ROC plots illustrate the diagnostic ability of the binary classifier in the segmentation model as the discrimination threshold varies by plotting the TPR (sensitivity or recall) against the FPR (1-specificity). a) Shows the class 0 (*OFF*) ROC curve with AUC value of 0.99. b) Visualizes the class-based comparison of ROC curves along with the macro-average (independently for each class) and micro-average (aggregative contribution for all classes) showing an AUC of 0.99 in both settings. Note that One-vs-One and One-vs-Rest class AUC has identical results given the 2-class problem in hand.

Chart

Description automatically generated

(a)

A picture containing graphical user interface

Description automatically generated

(b)

**Fig. S4 | Illustrating the trade-off between true positive rate and positive predictive value in different probability thresholds in segmentation.** a)Micro-averaged (aggregative contribution for all classes) precision-recall score (area under the curve) for both classes was 0.99. b) The precision-recall values were 0.99 and 0.98 for each class of forceps *ON* (class 1) and *OFF* (class 0), respectively. The inclusion of ISO-F1 curves shows that for all the points in the precision/recall space, most instances have F1 scores over 0.8 in the classification problem.

**Surgical Skill Classification**

Chart, line chart, histogram

Description automatically generated Chart, line chart

Description automatically generated

1. (b)

**Fig. S5 |** **Accuracy and loss function values during training and validation steps for force profile surgical skill pattern recognition using InceptionTime model.** History results for model accuracy and loss function value over 100 epochs overlaid for both training and validation iteration. a) The minimum validation loss function value occurred at epoch 23 and was 0.4760 (training loss = 0.4362). b) The historical accuracy for training has a consistent improvement over the training trials but becomes steady after around epoch 50 indicating an overfitting situation (achieved 0.98 in training and 0.68 in validation). To avoid overfitting, early stopping at epoch 23 was implemented.

Chart

Description automatically generated

**Fig. S6 | The summary prediction results for surgical skill prediction using InceptionTime model.** Across the 725 records of force segments with 51.5% in *Expert* labels and 48.5% in *Novice* labels, classification accuracy was 0.77, sensitivity (True Positive Rate: TPR) was 0.80, and specificity (True Negative Rate: TNR) was 0.73.

Chart, line chart

Description automatically generated

(a)

Chart, line chart

Description automatically generated

(b)

**Fig. S7 | The prediction performance for the skill classification model through receiver operating characteristic curves.** The ROC plots illustrate the diagnostic ability of the binary classifier in the skill classification model as the discrimination threshold varies by plotting the TPR (sensitivity or recall) against the FPR (1-specificity). a) Shows the class 0 (*Novice*) ROC curve with AUC value of 0.85. b) Visualizes the class-based comparison of ROC curves along with the macro-average (independently for each class) and micro-average (aggregative contribution for all classes) showing an AUC of 0.85 in both settings. Note that One-vs-One and One-vs-Rest class AUC has identical results given the 2-class problem in hand.

Chart, line chart, scatter chart

Description automatically generated

(a)

Chart

Description automatically generated

(b)

**Fig. S8 | Illustrating the trade-off between true positive rate and positive predictive value in different probability thresholds in the skill classification.** a)Micro-averaged (aggregative contribution for all classes) precision-recall score (area under the curve) for both classes were 0.85. b) The scores for each individual class of *Novice* (class 0) and *Expert* (class 1) were 0.82 and 0.87, respectively. The inclusion of ISO-F1 curves show that for all the points in the precision/recall space, almost half of the instances have F1 scores over 0.6 in the classification problem.

**Surgical Task Recognition**

A picture containing chart

Description automatically generatedChart

Description automatically generated

1. (b)

**Fig. S9 |** **Accuracy and loss function values during training and validation steps for force profile surgical task pattern recognition using InceptionTime model.** History results for model accuracy and loss function value over 150 epochs overlaid for both training and validation iteration. a) The minimum validation loss function value occurred at epoch 116 and was 0.0974 (training loss = 0.0120). b) The historical accuracy for training has a consistent improvement over the training trials but becomes steady after around epoch 50 indicating an overfitting situation (achieved 0.99 in training and 0.97 in validation). To avoid overfitting, early stopping at epoch 116 was implemented.

A picture containing table

Description automatically generated

**Fig. S10 | The summary prediction results for surgical task recognition using InceptionTime model.** Across the 2857 records of force segments with 15.05% in coagulation, 31.68% in pulling, 28.25% in manipulation, and 10.4% in dissecting, and 14.62% in retracting, the average accuracy of classification was 0.98, sensitivities (True Positive Rate: TPR) were 0.98, 0.98, 0.99, 0.96, and 0.98, respectively for the task classes.

Chart, line chart

Description automatically generated

(a)

Chart, line chart

Description automatically generated

(b)

**Fig. S11 | The prediction performance for the task recognition model through receiver operating characteristic curves.** The ROC plots illustrate the diagnostic ability of the classifier in the task classification model as the discrimination threshold varies by plotting the TPR (sensitivity or recall) against the FPR (1-specificity). a) Shows the class 0 (*Coagulation*) ROC curve with AUC value of 1.00. b) Visualizes the class-based comparison of ROC curves along with the macro-average One-vs-One and One-vs-Rest comparisons (independently for each class) as well as the micro-average values (aggregative contribution for all classes) which were equal to 1.00.

Chart

Description automatically generated

(a)

A picture containing graphical user interface

Description automatically generated

(b)

**Fig. S12 | Illustrating the trade-off between true positive rate and positive predictive value in different probability thresholds in the task recognition.** a)Micro-averaged (aggregative contribution for all classes) precision-recall score (area under the curve) for all classes were 1.00. b) The inclusion of ISO-F1 curves show that for all the points in the precision/recall space, most of the instances for all classes have F1 scores over 0.8 in the classification problem. The precision-recall score for each class was: *Coagulation* (class 0) = 0.99, *Pulling* (class 1) = 1.00, *Manipulation* (class 2) = 1.00, *Dissecting* (class 3) = 0.99, and *Retracting* (class 4) = 1.00.

**Methods**

* for each (Fr, idx &Fl, idx) in Fs:

if (MAidx (Fr, window=5) & MAidx (Fl, window=5) <= 0.3) & (init\_seg\_ididx = 0):

remove Fr, idx &Fl, idx

Fr, idx: idxth data point of the right prong force profile.

Fl, idx: idxth data point of the left prong force profile.

Fs: SmartForceps force profile time-series.

MAidx (*X*, window=*w*): Moving average of time-series *X* with a window size of *w* at idxth data point.

init\_seg\_ididx: Initial segmentation ID (i.e., 0: OFF, 1: ON) at idxth data point.

**Fig. S13 | Pseudo code for rule-based data point filtering to balance *ON* and *OFF* data samples.**

A rule-based algorithm was designed to remove the excessive inactive time points when the rolling average with a window of 5 for the left and right prong forces were less than or equal to 0.3 (N). Data points with overlapping 0 (*OFF*) labels in both rule-based and manually labelled indices were removed.



**Fig. S14 | The U-Net model architecture used for segmentation of force signals.** The graph shows detailed procedure names and attribute values for force profile segmentation model. The network comprised of different *convolutional* *encoder* and *decoder* operations with residual connection bypass and intermittent operations, e.g., *batch normalization*, *activation*, *max pooling*, and *concatenation layers*, and a final segment classifier with *activation*.The visualization was created in <https://netron.app>.



**Fig. S15 | The InceptionTime model architecture used force profile pattern recognition on time-series of segmented force profiles.** The graph shows detailed procedure names and attribute values for skill classification model (depth size = 8). Note: the network for task recognition is not included in the report to avoid duplication. The network included multiple layers including a stacked series of *convolutional layers* to learn the features followed by a concatenation *layer*, a *bottleneck layer* to reduce the dimensionality accompanied by a *max pooling layer*. The extracted features, as a new dimension, were fused into the network after *resampling* and *normalization*. as a new dimension to the network. The last layer shaped the probabilities of different classes, e.g., surgical proficiency scores or the task categories. The visualization was created in <https://netron.app>.



**Fig. S16 | The LSTM model architecture used as an experimental model for force profile pattern recognition on time-series of segmented force profiles.** The graph shows detailed procedure names and attribute values for task recognition model. Note: the network for skill classification recognition is not included in the report to avoid duplication. The network was comprised of an *LSTM layer* with *TanH activation* to interpret the extracted features, a *dropout regularization layer*, a *ReLU activation layer*, **and an output layer with *Softmax activation* providing the probability distribution of each surgical task class.** The visualization was created in <https://netron.app>.