**Methods**

**Force Profile Segmentation**

***Rule-Based Data Balancing for Segmentation***

A rule-based data point filtering was applied to mitigate the problem of imbalanced data in *On* and *OFF* conditions for the recorded force data. In fact, 93.7% of the force data points were labeled as OFF (among a total of 11.6 million records), meaning that inactive status constructs most operating room times for SmartForceps. The algorithm performed inactive state removal by eliminating the excessive idle time points when the rolling average with a window of 5 for the left and right prong forces was less than or equal to 0.3 (N). The points with overlapping *OFF* labels in both rule-based and manually labeled data were removed from the analysis (data size reduced to approximately 398K records) (Fig. S1). This data regularization method resulted in 54.4% in *ON* labels and 45.6% in *OFF* labels, making the segmentation labels balanced across the two classes.

* 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. S1 | 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.

***Model Characteristics***

A custom-designed U-Net (T-U-Net: Time-series-U-Net) model was implemented that consisted of a convolutional encoder and decoder structure to capture the properties and reconstruct the force profile (: fixed-length segment interval each containing data points through channels for left and right prong) through a deep stack of feature maps. A mean-pooling-based classifier follows this on point-wise confidence scores for interval-wise time-series segmentation (: final segment intervals containing2 segment classes, i.e., device ON/OFF). For the training parameters, we considered 50 epochs, *Adam* as the optimizer, *Categorical Cross-Entropy* as the loss function, and *accuracy* and *validation loss* as the evaluation metrics for a random 10% subset of training data as the validation data. Grid search was performed over the learning rate, i.e., within [0.0001-0.1], T-U-Net filter values, i.e., within [16-128], and batch size, i.e., within [32-128] for hyperparameter tuning. The compact visualization of the model architecture is provided in Fig. S2, and the expanded view of the model created by <https://netron.app> is present in Fig. S3.



Diagram

Description automatically generated

**Fig. S2 | Compact visualization of T-U-Net model architecture used for segmentation of force signals.** The network is comprised of different *convolutional* *encoder* and *decoder* operations with residual connection bypath and intermittent operations, e.g., *batch normalization*, *activation*, *max pooling*, *up sampling* and *concatenation layers*, and a final segment classifier with *spatial dropout* and *activation*.The visualization was created in Microsoft PowerPoint version 16.49 with the icons obtained from a Google search: e.g., <https://www.iconfinder.com>.



**Fig. S3 | The T-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>.

**Surgical Force Pattern Recognition**

***Task Recognition Data Augmentation***

To mitigate the imbalance data issue for the task recognition procedure, i.e., 80.6% coagulation, 7.4% pulling, 5.2% manipulation, 3.5% dissecting, and 3.3% retracting, a synthetic time-series generation technique based on dynamic time warping (DTW) and Stochastic Subgradient (SSG) averaging was implemented 1. In this technique, new time-series samples were generated for each class carrying their characteristics based on a modified k-means clustering algorithm, in which the clustering was applied separately for each class, and the cluster centroids were used as new data points for the respective classes. To assign data to each centroid, DTW, a point-to-point measure for similarity comparison between two temporal sequences of data which may vary in speed, was used as the distance measure, and to calculate new centroids, Schultz and Jain’s stochastic subgradient mean algorithm was used as the averaging mechanism. To obtain the desired augmented time-series, we used *k* = 1 (number of k-means iterations), *ssg\_epochs* = 1 (number of iterations for the SSG algorithm), *n\_base* = 2 (controls the number of centroids to be generated), and *n\_reps* = 10 (number of iterations the algorithm is called). The generated data in each iteration was fed back to the model to create new data. Following data augmentation, the class data distribution was updated to 28.2% coagulation, 27.2% pulling, 22.2% manipulation, 12.9% dissecting, and 9.5% retracting (Fig. S15).

***FTFIT Model***

A model based on InceptionTime was developed, i.e., FTFIT (Force Time-series Feature-based InceptionTime), where force time-series related features were merged as new dimensions to the network and the architecture was optimized to fit the specifics of tool-tissue interaction forces in surgery. The input to the network was a segmented force time-series (segment intervals over 2 channels of left and right prong data in SmartForceps)*.* The network included multiple layers including, a *bottleneck layer* to reduce the dimensionality after a *max-pooling layer*, a stacked series of *convolutional layers* to learn the features, followed by a *concatenation layer*. Finally, the extracted features were fused into the network after *resampling* and *normalization* as a new dimension to the network. The network's output was the probabilities of different classes, e.g., surgical proficiency scores or the task categories. As the evaluation metrics, we used *Adam* optimizer on ***Categorical Cross-Entropy*** along with a customized loss function (details in the supplementary codes), *accuracy* and *validation loss*. We applied grid search over the learning rate, i.e., within [0.001-0.1], and network depth, i.e., within [6-12] layers, input data window size, i.e., within [96-200], and batch size, i.e., within [32-128] for hyperparameter tuning. The compact visualization of the model architecture is provided in Fig. S4, and the expanded view of the model created by <https://netron.app> is present in Fig. S5.



**Diagram

Description automatically generated**

**Fig. S4 | Compact visualization of FTFIT model architecture used for classification of surgical skills and tasks using time-series of segmented force profiles.** 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*. As a new dimension, the extracted features were fused into the network after *resampling* and *normalization*. The last layer shaped the probabilities of different classes, e.g., surgical proficiency scores or the task categories.The visualization was created in Microsoft PowerPoint version 16.49 with the icons obtained from a Google search: e.g., <https://www.iconfinder.com>.



**Fig. S5 | The FTFIT 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>.

***LSTM Model***

A recurrent neural network based on LSTM that includes an input layer for the segmented force data (), *LSTM layers* 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 network weights which characterizes the behavior of transformations were identified through nonlinear optimization methods, i.e., *Adam*, to minimize the loss **function, e.g., *Categorical Cross-Entropy* and**a customized loss function (details in the supplementary codes)***,*** in the training data and backpropagation of error throughout the network for updating the weights.The performance of our models was evaluated by generalization through testing on previously unseen data using *accuracy* and *validation loss*. A grid search was applied over the learning rate (between 0.001-0.1), the LSTM unit size (between 100-600), input data window size (between 96-200), and batch size (between 32-128) to tune the hyperparameters. The model architecture visualization is shown in Fig. S6.





**Fig. S6 | 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>.

**Table S1: List and Description of Hand-crafted Features for the Surgical Pattern Recognition Models**

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Duration Force | Duration of force application in one task segment |
| Mean Force | Average of force values in one task segment |
| Maximum Force | Maximum of force values in one task segment |
| Minimum Force | Minimum of force values in one task segment |
| Range Force | Range of force values in one task segment |
| Median Force | Median of force values in one task segment |
| Force Standard Deviation | Standard deviation of force values in one task segment |
| Force Coefficient of Variance | Coefficient of variation of force values in one task segment |
| Force Distribution Skewness | The extent to which the force data distribution deviates from a normal distribution |
| Force Distribution Kurtosis | The extent to which the force data distribution is tailed in a normal distribution |
| Force Distribution Normality Test | Shapiro-Wilk test of normality in force data distribution |
| Force Profile Peaks Count | Number of force peaks in one task segment |
| Force Profile Maximum Peak Value | Force peak maximum value in one task segment |
| Force Time Series Frequency | Dominant time-series harmonics extracted from Fast Fourier Transform (FFT) of force value in one task segment |
| Force Time Series Period Length | Average time length of force cycles in one task segment |
| Force 1st Derivative Standard Deviation | Standard deviation for the first derivative of the force signal in one task segment |
| Force Profile Flat Spots | Maximum run length for each section of force time-series when divided into ten equal-sized intervals |
| Force Profile Trend Strength | Force time-series trend in one task segment |
| Force Profile Linearity | Force time-series linearity index (from Teräsvirta’s nonlinearity test) in one task segment |
| Force Profile Stability | Force time-series stability index (variance of the means) in one task segment |
| Force Profile Lumpiness | Force time-series lumpiness index (variance of the variances) in one task segment |
| Force Profile Crossing Points | Number of zero crossings in in one task segment |
| Force Profile Entropy | Force time-series forecastability in one task segment (low values indicate a high signal-to-noise ratio) |
| Force Profile Heterogeneity | Force time-series heterogeneity in one task segment (based on autoregressive conditional heteroskedasticity (ARCH) effects) |
| Force Profile Spikiness | Force time series spikiness index (variance of the leave-one-out variances of the remainder component) in one task segment |
| Force Profile First Autocorrelation Minimum | Time of first minimum of the autocorrelation function in force time-series signal from one task segment |
| Force Profile First Autocorrelation Zero | Time of first zero crossing of the autocorrelation function in force time-series signal from one task segment |
| Autocorrelation Function E1 | First autocorrelation coefficient from force time-series signal in one task segment |
| Autocorrelation Function E5 | Sum of the first ten squared autocorrelation coefficients from force time-series signal in one task segment |

**Results**

**Force Profile Segmentation**

Chart

Description automatically generatedA picture containing chart

Description automatically generated

1. (b)

**Fig. S7 |** **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. S8 | 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. S9 | 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. S10 | 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. S11 |** **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. S12 | 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. S13 | 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. S14 | 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**

Chart

Description automatically generated

**Fig. S15 | Scatterplot with a histogram of original force profile segment data vs. augmented samples for minority classes of surgical tasks.** The figure shows the distribution plot for Range vs. Mean force values of the resampled (200 data points for each prong) segment force profiles in the original (green) and augmented (red) samples. To have a detailed view, each surgical task has been differentiated using a specific circular size (*Coagulation* having the smallest size). The boxplots of each distribution have also been created, showing that the augmentation properly follows similar characteristics of the original data samples, i.e., Mean (SD) Range = 1.04 (0.76) (N) and Mean (SD) Mean = 0.23 (0.39) (N).

A picture containing chart

Description automatically generatedChart

Description automatically generated

1. (b)

**Fig. S16 |** **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. S17 | 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. S18 | 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. S19 | 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.

**References**

1. Schultz D, Jain B. Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. Pattern Recognition 2018;74:340-58.