Pattern-Based Behavioural Analysis on Neurosurgical Simulation Data

Scott Buffett SCOTT.BUFFETT@NRC.GC.CA
Research Centre for Digital Technologies
National Research Council Canada
Fredericton, New Brunswick, Canada

Catherine Pagiatakis CATHERINE.PAGIATAKIS@NRC.GC.CA
Research Centre for Medical Devices
National Research Council Canada
Boucherville, Quebec, Canada

Di Jiang DI.JIANG@NRC.GC.CA
Research Centre for Medical Devices
National Research Council Canada
Boucherville, Quebec, Canada

Abstract

This paper presents the results of an analytics-based study to determine key differences in junior resident level and expert level surgical skill when engaging with a neurosurgical simulator. Window-based time series discretization and sequential pattern analysis were used on positional data to identify frequent movement patterns and instrumentation techniques associated with each skill class. Cross-validation confirmed that a Bayesian classification model constructed using these patterns can be used to predict skill level with a high degree of confidence and accuracy on a small sample of neurosurgeons who engaged with the simulator. An analysis of movement speed also revealed that the junior residents exhibited a high degree of very slow and very fast movements, whereas the expert surgeons displayed a significantly more consistent technique of moderate-speed movements. Finally, the analysis was integrated within a cloud-based learning framework, helping to provide beneficial feedback on movement proficiency to resident surgeons in training. The presented work makes two key contributions to the field of machine learning in the medical field: the study 1) employs a low-level behaviour-based analysis of surgical technique, as opposed to high-level summary metrics such as blood loss and average force, and 2) avoids the use of expert information on neurosurgical skill within the AI engine, and thus employs an analysis that is entirely uninformed.

1. Introduction

Platforms and devices that support the progress and development of learners in a particular domain or area of expertise have existed for quite some time in the form of software simulators, intelligent tutors and online courseware. In the medical domain, suture practice kits and cadavers, and more recently training simulators such as SIM*VIVO[1], Touch Surgery[2]

[1] https://sim-vivo.com/
[2] https://www.touchsurgery.com/

and NeuroTouch[3] have facilitated young medical residents to practice their technique outside of the operating room, as well as experienced surgeons to prepare for complicated procedures, or even to hone their skills as they learn to handle new procedures or implement new techniques.

Where trends in the learning domain have seen a recent shift is in the area of learning analytics, particularly due to both an increasing ability to generate and collect data from learner interactions with these devices and platforms, and the rise in analytical technologies that can facilitate the analysis of this data for the extraction of key information that can be used to further enhance or accelerate the learning experience.

Technical Significance This work demonstrates the use of sequential pattern mining techniques to identify key movements that differentiate surgical skill, which are then validated via the performance of a classification model constructed using these movements. The paper makes two key contributions to the field of machine learning in healthcare, specifically in what concerns the application of machine learning techniques in medical education: the study 1) employs a low-level behaviour-based analysis of surgical technique, as opposed to high-level summary metrics such as blood loss and average force, and 2) avoids the use of expert information on neurosurgical skill within the AI engine, and thus employs an analysis that is entirely uninformed on any baseline behaviour of what might constitute skilled or unskilled technique.

Clinical Relevance Techniques proposed in this paper offer significant advances in the area of simulation-based surgical training, by facilitating the ability to offer direct, personalized recommendations for improvements in surgical technique with emphasis on skills derived from a low-level analysis, that may not be observable at high-level metrics perspective. This will result in more targeted and effective feedback for a learner and thus faster mastery of expert skills and manipulation.

1.1. Neurosurgical Simulation

The National Research Council Canada (NRC), in collaboration with teaching hospitals from across Canada, developed NeuroTouch, a virtual reality simulator that is currently being used at 17 teaching hospital locations worldwide, including Canada, U.S., Europe, Africa and Asia, to train surgical residents in endoscopic and craniotomy-based neurosurgical oncology procedures. It provides a risk-free training environment that can reduce medical errors and improve patient safety, all while decreasing operating room costs that are incurred through traditional training methods. The simulator is equipped with two haptic devices with 3-degrees of freedom that allow tactile-based interaction between a surgical instrument and a virtual soft tissue. Simulation of tissue behaviour resulting from tissue-tool interaction, as well as bleeding dynamics, are physics-based, which give rise to realistic training scenarios. The NeuroTouch simulator has more than 30 simulation exercises that target training of different skills, ranging from fundamental instrument handling exercises to tumour resection.

[3] Commercially available under the name NeuroVR (CAE Healthcare, https://caehealthcare.com/surgical-simulation/neurovr)

One of the benefits of virtual reality training simulation is the capacity to monitor and record performance metrics. Specifically, for each simulation scenario, the NeuroTouch simulator records time series data for numerous scenario parameters (bleeding measures, tissue volumes etc.), surgical instrument data (spatial and handling parameters), and tissue-tool interaction information (contact with the tissue and duration of excessive force applied on the tissues). By using this data, scenario-specific performance metrics and overall scores were developed and corroborated by an international consortium of expert neurosurgeons.

1.2. Related Work

Machine learning pattern-based techniques have recently gained traction in the medical domain, particularly in the area of simulation-based training. Kennedy et al. (2013) analyzed student interactions with a temporal bone surgical simulator, in an effort to provide real-time feedback on surgical technique with regard to the level of skill being exhibited. In this case, high-level discrete metrics were observed (e.g. whether force being applied falls above or below some threshold), and Hidden Markov Models were used to predict skill level, where low-level instrument manipulation (e.g. a "stabbing" motion) was represented and implied by the unobservable or "hidden" states. This work built on that of Sewell et al. (2008), who employed Naïve Bayes classification on high-level features such as average force, suction position, etc., collected from mastoidectomy simulation to classify expert and novice proficiency. Toussaint and Luengo (2015) applied sequential rule mining techniques to further analyze instrumentation technique in vertebroplasty simulations.

Several studies based out of the Montreal Neurological Institute (MNI) have been previously executed using the NeuroTouch simulator so as to define and validate performance metrics and proficiency performance benchmarks related to psychomotor skills and neurosurgical competency (Sawaya et al., 2017; Gélinas-Phaneuf et al., 2014; Azarnoush et al., 2017; Winkler-Schwartz et al., 2016; Azarnoush et al., 2015; Bugdadi et al., 2018). In these studies, a set of metrics assessing safety, quality and efficiency were proposed and evaluated. The metrics were classified as Tier 1 (extracted directly from NeuroTouch), Tier 2 (derived from outputted NeuroTouch data) and Tier 3 (advanced metrics derived from outputted NeuroTouch data). Several of these investigations found statistically-significant differences between the performance of board-certified expert neurosurgeons and neurosurgical residents, in particular, for metrics concerning safety and efficiency. These studies did not explore baseline behaviour that could be representative of skilled or unskilled technique based on an uninformed analysis, but rather validated a set of predefined metrics that were proposed with direct input from experts.

Overall, one key shortcoming of the current state of the art and corresponding learning analytics systems is that they constitute an outcome-based approach where feedback is typically given at a high-level, such as scores or measurements for various skills, modules, aspects of the course, etc. While this can lend information to the trainee on what areas need improvement and thus where further practice or study is required, it does not indicate exactly how the trainee's behaviour led to these outcomes. On top of facilitating recommendation for future study, a low-level behaviour-driven approach can present the learner with feedback that can be used immediately, possibly with the assistance of an instructor, to rectify deficiencies in the learner's technique.

1.3. Objectives

In this study, machine learning techniques using sequential pattern mining were implemented to conduct a pilot behavioural analysis on a dataset obtained from NeuroTouch simulator. Data was collected from a number of physicians in the neurosurgical field who self-reported their skill class as neurosurgical residents (junior: Post Graduate Year, PGY 1-3; senior: PGY 4-6) or expert (board-certified) surgeons. The first objective of the analysis was to identify patterns in low-level behaviour pertaining to how the subjects manipulated the instruments, that could differentiate between the different skill classes, by using an uninformed analysis that does not depend on domain-specific input. Specifically answers to two questions were sought as part of this first objective: 1) how do common patterns generally differ between different skill classes, and 2) what are some specific examples of common patterns that are attributable to each skill class? Contingent on the successful identification of patterns indicative of surgical skill level, the second objective was to present three types of information to a learner: 1) scores determined based on the proficiency of instrument manipulation, 2) graphical representations of a learner's common patterns that are a) deemed skillful and b) deemed unskillful, and 3) recommendations for how a learner's technique could be improved.

2. Methodology

2.1. Neurosurgical Simulation Dataset

The NeuroTouch simulation performance data used in this study was collected during the 2015 Annual Congress of Neurological Surgeons (New Orleans, USA), using a case study format.

2.1.1. Study Participants

Participation in the study was on a volunteer basis and all collected simulation results and participant demographic data was anonymized. Each participant was presented with a consent form and an information sheet describing the goal of the case study and instructions for the simulation exercise to be executed. Participants were asked to self-identify their skill class classification as either a neurosurgical resident () or a staff (expert) neurosurgeon. Demographic data collected for each of the participants included age, gender, handedness, level of training (Post Graduate Year, in the case of residents), years in practice (in the case of staff surgeons), approximate number of meningioma cases performed, approximate number of endoscopic third ventriculostomy cases performed, the average number of hours per week spent playing a musical instrument and the average number of hours per week spent playing video games. All data was collected with approval for corresponding governing body ethics review boards.

2.1.2. Simulation Scenario

Considering the context of the data collection as well as the resulting pool of candidates for the study cohort, a tumour debulking exercise on a synthetically-generated tumour was selected for this study as the target simulation scenario (Tumour Debulking 101 scenario

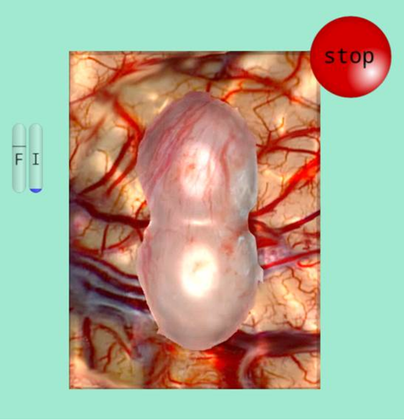


Figure 1: Tumour Debulking simulation scenario utilized in the case study.

Figure 1). This scenario allowed to minimize the required participation time, ensure a feasible level of difficulty and target the required skill set.

Participants were instructed to debulk the meningioma-like inclusion and to clear the field of blood, ensuring that bleeding has stopped completely, all while avoiding to remove surrounding healthy brain tissue and avoiding to push forcefully on the tissues. The selected scenario required bimanual manipulation of the surgical instruments, with a suction instrument used in the non-dominant hand to remove blood from the scene and a cavitron ultrasonic aspirator (CUSA) used in the dominant hand to aspirate the tumour.

For each study participant, the following data was collected from the simulator:

1. summary file containing the metrics utilized in the computation of the overall performance score for the scenario

2. time-series data, collected at a frequency of approximately 50Hz containing all the simulation and scene status data

3. final screen shot of the simulation scene prior to scenario termination

2.1.3. Data Preprocessing

In order to ensure the quality of the data, several iterations of preprocessing were carried out on the collected data:

1. Manual verification on the participant's skill class: a comparison was made between the self-reported skill class of the participant and the total number of practice years/level of training and/or the total number of surgical cases performed. Neurosurgical residents were classified as junior (PGY 1-3) or senior (PGY 4-6).

2. Manual performance estimate of the participant's skill class: a rough estimate of performance level was made based on the final screen shot of the simulation scene and used as an extra tag for the designated skill class of the user.

3. Removal of low-performance data: study coordinator notes were used to filter out portions of the logged raw performance data that were caused by disruptions and/or interruptions.

The outcome of the preprocessing step was a clean, labeled data set together with a brief description of each performance result that was ready for use in the machine learning pattern analysis step presented in subsequent sections.

2.2. Pattern Analysis with Neurosurgical Simulation Data

Contemporary research in frequent pattern mining research can be traced back to the seminal work of Agrawal and Srikant (1994, 1995) for itemset mining and sequential pattern mining with the development of the *Apriori* and *AprioriAll* algorithms, respectively. Sequential pattern mining work was further advanced with performance-improving algorithms such as *GSP* (Srikant and Agrawal, 1996), *SPADE* (Zaki, 2001), *SPAM* (Ayres et al., 2002) and *PrefixSpan* (Pei et al., 2001), and also with specialized algorithms for closed (Gomariz et al., 2013), high utility (Yin et al., 2012) and incremental (Wang and Tan, 1996) sequential pattern mining.

Due to the discrete data requirements of existing algorithms that seek to identify identically recurring constructs, frequency-based pattern mining and classification of movements in continuous space requires special attention. Alwasel et al. (2017) examined common movement patterns exhibited by concrete masons that are representative of skilled or safe execution, by using k-means clustering to determine a set of key poses. Each position in 3D space was then considered identical to the pose representing that position's nearest centroid. Das et al. (1998) addressed a further complexity to this problem by clustering subsequences or "windows" of time series data, which can be used to discretize sequences of positions, rendering the ability to consider frequently occurring movements in continuous space.

2.2.1. Sequential Pattern Mining

Sequential pattern mining (SPM) (Agrawal and Srikant, 1995; Mooney and Roddick, 2013) is a research discipline within the field of data mining that focuses on identifying frequently occurring sequences of objects or events. Let I be a set of *items*, and S be a set of *input sequences*, where each $s \in S$ consists of an ordered list of *itemsets*, or sets of items from I, also known as *transactions*. A sequence $\langle a_1 a_2 \ldots a_n \rangle$ is said to be *contained* in another sequence $\langle b_1 b_2 \ldots b_m \rangle$ if there exist integers $i_1, i_2, \ldots, i_n$ with $i_1 < i_2 < \ldots < i_n$ such that $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots, a_n \subseteq b_{i_n}$. A sequence $s \in S$ *supports* a sequence s' if s' is contained in s. The support $sup(s')$ for a sequence s' given a set S of input sequences is the percentage of sequences in S that support s', and is equal to $sup(s') = |\{s \in S | s \text{ supports } s'\}| / |S|$. A sequence s' is deemed a *sequential pattern* if $sup(s')$ is greater than some pre-specified minimum support. Such a pattern with a total cardinality of its itemsets summing to n is referred to as an *n-sequence* or *n-pattern*. A sequential pattern s' is a *maximal sequential pattern* in a set S' of sequential patterns if $\forall s'' \in S'$ where $s'' \neq s'$, s'' does not contain s'. The general goal of sequential pattern mining is then to identify the set S' that contains all (and only those) sequences that are deemed sequential patterns according to the above. In some cases, the set consisting of only maximal sequential patterns is preferred.

Sequence Database	1-seq	2-seq	3-seq
$\langle \{b, c\}, \{c, d\}, \{e\} \rangle$	$\langle \{a\} \rangle$	$\langle \{a\}, \{d\} \rangle$	$\langle \{b, c\}, \{d\} \rangle$
$\langle \{a, c\}, \{b, c\}, \{d\} \rangle$	$\langle \{b\} \rangle$	$\langle \{b, c\} \rangle$	$\langle \{b\}, \{d\}, \{e\} \rangle$
$\langle \{c\}, \{e\} \rangle$	$\langle \{c\} \rangle$	$\langle \{b\}, \{d\} \rangle$	$\langle \{c\}, \{d\}, \{e\} \rangle$
$\langle \{c\}, \{d\}, \{e, f\} \rangle$	$\langle \{d\} \rangle$	$\langle \{b\}, \{e\} \rangle$	
$\langle \{a, b\}, \{d\}, \{e\} \rangle$	$\langle \{e\} \rangle$	$\langle \{c\}, \{c\} \rangle$	
		$\langle \{c\}, \{d\} \rangle$	
		$\langle \{c\}, \{e\} \rangle$	
		$\langle \{d\}, \{e\} \rangle$	

Figure 2: Example sequence database with mined sequential patterns using a minimum support of 0.4.

To illustrate, consider the example set S of sequences given in the first column of Figure 2. Using a minimum support of 0.4, the set of all sequential patterns is indicated, specified by 1-sequences, 2-sequences and 3-sequences.

Note that, while this table gives all of the frequent sequential patterns, only $\langle \{a\}, \{d\} \rangle$, $\langle \{c\}, \{c\} \rangle$, $\langle \{b, c\}, \{d\} \rangle$, $\langle \{b\}, \{d\}, \{e\} \rangle$ and $\langle \{c\}, \{d\}, \{e\} \rangle$ are *maximal* sequential patterns.

2.2.2. Sequence Classification

Sequence classification (Lesh et al., 1999, 2000; Xing et al., 2010)) is the field of study that attempts to classify sequences in S by using frequent sequential patterns as features in the classification. Consider the above model with the addition of a set L of *class labels*, where each $s \in S$ is labeled with an element of L. S is now a set of *examples*, where each example $s \in S$ can be represented by a set of feature-value pairs using features from the set S' of frequent sequential patterns and boolean values. A feature f is thus assigned the value "true" if s contains f, and "false" otherwise. For example, consider features $f_1 = \langle \{b\}\{c\}\{d\} \rangle, f_2 = \langle \{b\}\{d\}\{e\} \rangle, f_3 = \langle \{c\}\{d\}\{e\} \rangle$. The sequence $\langle \{a, b\}\{d\}\{e\} \rangle$ could then be represented by the feature-value pairs $(f_1, false), (f_2, true), (f_3, false)$.

The goal of sequence classification is to identify sequences for the feature set that have the following properties:

- Features should be frequent
- Features should be distinctive of at least one class
- Feature sets should not contain redundant features

The typical methodology for sequence classification involves two phases: 1) sequential pattern mining to discover the frequent sequences that potentially will make good feature candidates, and 2) feature assessment (e.g. using an "interestingness" measure) to select the best candidate. From there, standard machine learning based classification methods such as SVM or Naïve Bayes can be used to build a classification model and label future instances.

2.2.3. Continuous Input Values

The sequential pattern analysis algorithms incorporated in this paper inherently rely on discrete-valued datasets. However, the positional data used in this study was represented by real-valued coordinates in three-dimensional space. Therefore a discretization technique from the literature that is designed to convert time series data into discrete sequences for the purpose of extracting information in the form of rules (Das et al., 1998), was used.

The technique works by taking small intervals, or *windows* of consecutive data, and mapping these windows to discrete elements. Thus, since a discrete element refers to a series of values, in the context of this work the element can be referred to as a "movement". Windows that are sufficiently similar are mapped to the same element. Tolerance for a slight variation in the data mapped to the same discrete element allows for the identification of frequent patterns of "similar" activity in the data. Due to the fact that the data is continuous, searching for exact matches is likely to result in a high number of uninteresting patterns that happen very infrequently, likely occurring only once.

To further facilitate identification of frequent patterns, positioning is not considered to be absolute, but rather relative to the other data in the window. Thus an instrument's movement along the x-axis along $x = 4, 3, 6$ would be considered (exactly) similar to a movement along $x = 6, 5, 8$, since the relative movement between each point is the same.

2.2.4. Identification of Surgical Instrumentation and Movement Patterns

In this section the methodology used for identifying patterns in the data that are indicative of different skill classes is described. In the initial phase, data preprocessing is conducted to obtain the desired task-specific data. In particular, only the positioning of the CUSA instrument is collected, specifically when it is in use by the dominant hand and is in direct contact with the tumour. Data discretization is then employed as described in Section 2.2.3 to transform the continuous movement sequences associated with each skill class to a set of discrete movement sequences.

Each set of discrete movement sequences associated with a skill class is then mined individually for sequential patterns. For this step the SPAM (Ayres et al., 2002) algorithm is employed. The set of sequential patterns returned by the SPAM algorithm for each skill class thus offers the set of potential *features* for that skill class. Feature selection is then conducted by determining whether the feature pattern's support in the skill class is significantly higher than its support in any other skill class using a chi-squared test (Lesh et al., 1999). This process results in the behaviours that can then be displayed to the user as being indicative of the skill class from which they originated. Finally, a Naïve Bayes classifier is trained to model each skill class and predict the skill level exhibited in a candidate simulation. Leave-one-out cross-validation is used to assess the classifier's performance, which in turn lends insight into the likelihood of the selected patterns truly being representative of their corresponding skill classes.

3. Sequential Pattern Mining-Based Skill Proficiency Results

3.1. Experimental Setup

To test the ability to build a model of expert behaviour using the collected dataset described in Section 2.1, the procedure outlined in Section 2.2.4 was used. As a preparation step for the mining procedure, each dataset was divided into a number of sub-datasets, each consisting of 150 consecutive records. This allowed the model to consider behaviours that are done frequently and with regularity over 150-record segments, as opposed to those that only appear once or very few times over the course of the scenario. Following the testing of different groupings of skill classes, as outlined in Section 2.2.4, the final dataset used to train the classification model consisted of three junior residents and three experts. Data pertaining to senior residents was excluded from the data used to train the model.

Each record in the time-series data (captured at a frequency of approximately 50Hz) gives a snapshot of the positioning of the instruments used at a given time point. In this analysis, only the x, y and z positions of the tip of the CUSA were used in the model.

To facilitate an unbiased evaluation of the modeling capabilities, cross-validation was employed by leaving out all 150-record segments for a candidate participant, training the model on the remaining segments and then testing the model on the candidate dataset excluded from the training phase. A second level of testing was executed on the dataset pertaining to the senior residents (which were excluded from the final training model). In this testing phase, the hypothesis was that the senior residents should be classified into the "expert" group. The trained model used in this last testing phase used all junior and expert data.

3.2. Movement Analysis and Skill Level Prediction

Results of model performance related to the cross-validation experiment (presented in Section 3.1) are depicted in Table 1. Data processing, as described in previous sections, resulted in 354 junior segments and 146 expert segments. Segment accuracies are listed per user in the table, indicating how many of the user's segments were classified as "Junior" or "Expert".

If the final prediction for a user's skill class was made based on the relative frequency of junior-type to expert-type movements, then the predictions would have been correct in all cases except for that of expert1. However, employing a null hypothesis that the number of junior/expert segments are equal (and there is no significance in the results), expert1's p-value of 0.1 would have in fact resulted in failure to reject when employing a standard minimum threshold of 0.05. Thus it would not have been possible to (erroneously) predict with confidence that expert1 was a junior surgeon.

The results related to the model's skill class predictive accuracy of senior residents are displayed in Table 2. The hypothesis was that the data of the senior residents would be classified as expert level. Here, three of the five senior residents were correctly classified as "expert". However, a closer inspection shows that in the two incorrect cases, it was again not possible to reject the null hypothesis.

Skill Class	No. Segments Junior	No. Segments Expert	Accuracy	Prediction	p-value	Reject Null?
junior1	31	17	65%	Junior	0.03	y
junior2	52	10	84%	Junior	< 0.001	y
junior3	167	77	68%	Junior	< 0.001	y
expert1	37	26	41%	Junior	0.1	n
expert2	8	33	80%	Expert	< 0.001	y
expert3	10	32	76%	Expert	< 0.001	y

Table 1: Accuracy of skill prediction on junior and expert surgeons

Overall, out of 11 total test cases, in eight cases it was possible to reject the null hypothesis that there was no significance, with the surgeons' true level of skill predicted correctly in all eight of those cases.

Skill	Junior	Expert	Accuracy	Prediction	p-value	Reject Null?
senior1	14	37	73%	Expert	< 0.001	y
senior2	5	59	92%	Expert	< 0.001	y
senior3	25	19	43%	Junior	0.23	n
senior4	10	36	78%	Expert	< 0.001	y
senior5	17	15	47%	Junior	0.43	n

Table 2: Accuracy of skill prediction on senior surgeons

3.3. Movement Speed Analysis

While conducting our analysis of expert level patterns and further examining the differences between them and those exhibited by the junior residents, a prominent trend in the speed of movements with the CUSA, particularly in the z-axis (vertical movements), was observed. Specifically, junior residents were found to be much more likely to exhibit a larger percentage of both very slow and very fast vertical movements, while expert neurosurgeons were much more likely to exhibit more consistent, medium-level movement speeds. Further investigation confirmed these observations. Figure 3 depicts the percentage of movements at various speeds exhibited by both junior residents and expert surgeons. From the leftmost chart, it is easily apparent that a significantly higher percentage of junior movements are executed at a vertical speed less than 0.89 mm/s, where a higher percentage of expert movements are executed at vertical speeds higher than 0.89 mm/s but less than 5.95 mm/s. With percentages for both skill classes being close to zero for higher speeds, a second graph (on right) is provided at a finer level of granularity. Here it can more easily observed that the junior residents are reaching these high vertical speeds two to three (or more) times more frequently than their expert counterparts. One can surmise here that the junior residents may exhibit more of an up/down stabbing motion when coming in contact with the tumour.

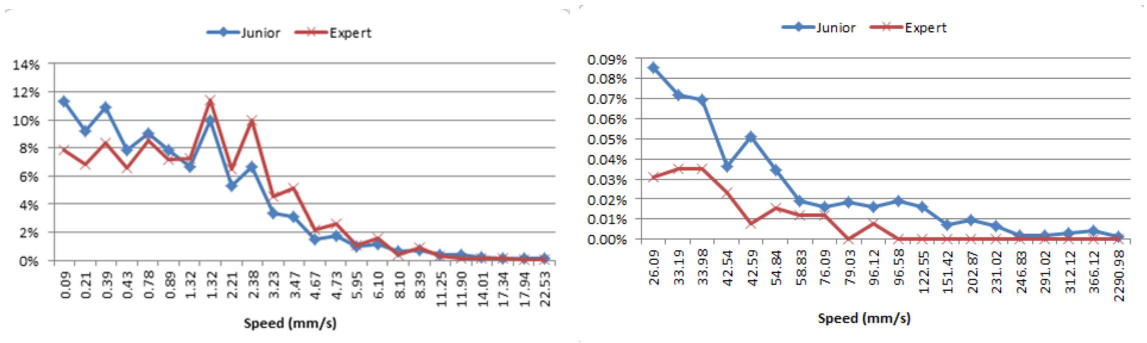


Figure 3: Percentage of movements at various speeds exhibited by junior vs expert surgeons

4. Analysis Integration Within Cloud-Based Framework

4.1. Direct Skill-Proficiency Mapping Framework

One of the limitations of the generic performance scoring system for the NeuroTouch simulation scenarios is that it is based on average metric values and is therefore deficient in indicating user proficiency. The overall score, while developed and corroborated by a consortium of expert neurosurgeons, is computed based on scenario-specific goals and errors and is presented in a much gamified method. The time-series simulation data, while available to the trainees, provides an overwhelming amount of data that proves to be of no direct help to them. Moreover, there exists an overlap of skills and competencies between the different simulation scenarios, but is not readily observable and exploitable based on the current training module structure of evaluation and performance. Finally, NeuroTouch does not provide recommendations based on user performance.

Therefore, in the context of the cloud-based personalized learning platform, a direct skill-proficiency mapping framework was developed such that user performance could be generalized and learning could be enhanced. As such, each NeuroTouch scenario related to craniotomy-based procedures was decomposed into a set of basic skills, for example, *clear blood* and *aspirate tumour*, which were compliant with targeted technical skills defined by the general curriculum for neurosurgical resident training. Subsequently, each skill was decomposed into a set of possible subskills, that define the quality, safety and efficiency with which each skill was executed. The use of a small subset of generalized skills and subskills permits the user to create links between their performance within each scenario, all while creating well-defined global and common competency goals.

Six proficiency levels were defined to describe user performance relevant to each skill and subskill, as can be seen in Table 3. For each craniotomy-based scenario, the skill and subskill proficiency definition was computed based on the scenario parameters and the corresponding grading scheme defined and corroborated by the expert neurosurgeon consortium.

4.2. The *Rabaska* Learning Platform

Personalized learning has recently been a leading trend in the education field. A remote performance visualization and learing platform, *Rabaska*, that permits users (including trainees,

Proficiency Level	Classification	Grade [%]
1	Insufficient	0 - 50
2	Mediocre	51 - 60
3	Satisfactory	61 - 80
4	Good	81 - 90
5	Excellent	91 - 95
6	Outstanding	96 - 100

Table 3: Proficiency level definition

surgeons, researchers and others) to access and visualize their simulation results anywhere and anytime, was developed. Rabaska features:

1. secured and generic data transfer from any registered NeuroTouch simulator device to and storage within a centralized database

2. centered- and simulation-specific performance result visualization that mimics the interface of the NeuroTouch simulator

3. performance evaluation based on the skill-proficiency framework presented in Section 4.1

4. personalized recommendation provision based on proficiency level

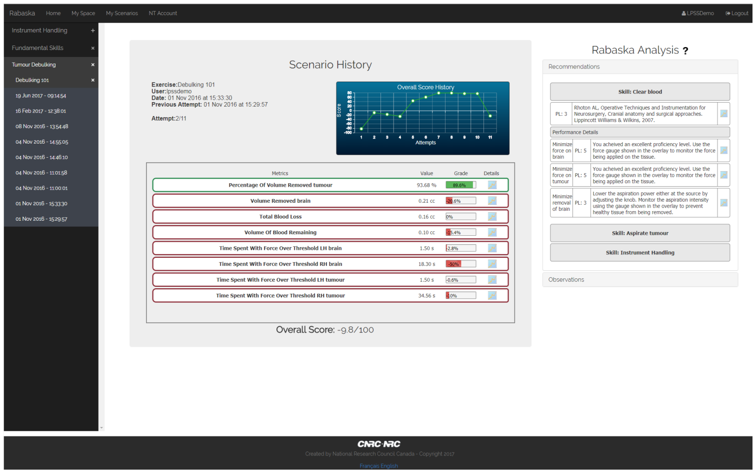
Rabaska enables the enhancement of a training experience by not only applying the proposed skill-proficiency framework, but also by providing personalized skill-oriented recommendations in order to improve performance. In addition, by storing all user simulation result data in a centralized location, it can be readily used in analytics-based pedagogical studies of performance. These studies allow for the extraction of performance indicators and metrics and detailed behaviour-related observations in what concern fine motor skills and instrument handling. Such insights could not only help a trainee achieve high levels of competency in an efficient manner, but also provide benchmark competency measures required for fulfillment of curriculum-defined technical skills requirements and board certification. In addition, results from these studies can subsequently be translated into improvement recommendations which promote personalized learning and further aid in the achievement of competency prior to the execution of surgical interventions on patients.

4.3. Integration of SPM Model Within Rabaska

Rabaska was developed using a Service-Oriented Architecture (SOA), which permitted the SPM model to be integrated seamlessly within the framework, all while functioning as an independent service module. There is a bidirectional communication between it and the SPM model via a pre-defined standard communication protocol. More specifically, for registered Rabaska users, each simulation result of the Tumour Debulking 101 scenario within NeuroTouch is sent to the SPM module upon upload to the centralized database of Rabaska, where it is tested against the trained SPM model. The trained model is used to classify common movements and instrument movement speed executed by the user as either junior- or expert-type and provides a corresponding proficiency level. This result is subsequently communicated back to Rabaska and utilized explicitly in the recommender component (entitled *Rabaska Analysis*) of the results. In particular, recommendations are provided to the user on how to improve their instrument handling so as to better reproduce movements excuted by expert neurosurgeons.

4.4. Performance Visualization in Rabaska

Figure 4 displays the performance visualization component of Rabaska. On the right-hand side of the view, the integrated skill-proficiency framework and corresponding recommendation provision module can be observed (*Rabaska Analysis*). Recommendation provision is presented in two categories, namely *Recommendations* and *Observations* which are detailed below.



Scenario History

Exercise: Debulking 101
User:lpssdemo
Date: 01 Nov 2016 at 15:33:30
Previous Attempt: 01 Nov 2016 at 15:29:57

Attempt:2/11

Metrics	Value	Grade	Details
Percentage Of Volume Removed tumour	93.68 %	89.6%	
Volume Removed brain	0.21 cc	-20.6%	
Total Blood Loss	0.16 cc	0%	
Volume Of Blood Remaining	0.10 cc	-15.4%	
Time Spent With Force Over Threshold LH brain	1.50 s	-2.8%	
Time Spent With Force Over Threshold RH brain	18.30 s	-50%	
Time Spent With Force Over Threshold LH tumour	1.50 s	-0.6%	
Time Spent With Force Over Threshold RH tumour	34.56 s	-10%	

Overall Score: -9.8/100

Rabaska Analysis ?

Recommendations

Skill: Clear blood

PL: 3	Rhoton AL, Operative Techniques and Instrumentation for Neurosurgery, Cranial anatomy and surgical approaches. Lippincott Williams & Wilkins, 2007.	

Performance Details

Minimize force on brain	PL: 5	You acheived an excellent proficiency level. Use the force gauge shown in the overlay to monitor the force being applied on the tissue.	
Minimize force on tumour	PL: 5	You acheived an excellent proficiency level. Use the force gauge shown in the overlay to monitor the force being applied on the tissue.	
Minimize removal of brain	PL: 3	Lower the aspiration power either at the source by adjusting the knob. Monitor the aspiration intensity using the gauge shown in the overlay to prevent healthy tissue from being removed.	

Skill: Aspirate tumour

Skill: Instrument Handling

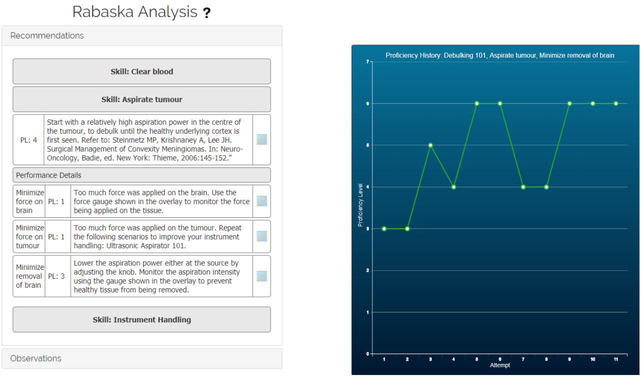
Observations

Created by National Research Council Canada - Copyright 2017

Figure 4: Rabaska: remote performance visualization platform with the integrated recommendation system.

The proposed sequential pattern analysis-driven recommendation system was integrated into *Rabaska Analysis*, as outlined in Section 4.3. Results and recommendations relating to movement speed, as described in Section 3.3, are presented under the *Recommendations* portion of the analysis (which also displays the pre-defined skills and subskill structure outlined in Section 4.1), and are categorized as a new skill, designated as *Instrument Handling*. Results pertaining to the common movements, as described in Section 3.2, are presented under the *Observations* portion of the analysis.

Figure 5 displays a more detailed view of the output of the direct skill-proficiency-recommendation module, with one particular skill, namely *aspirate tumour*, presented. This figure highlights the user-centric design of the recommendation module of Rabaska. For each skill related to the particular simulation scenario, the proficiency level is provided along side the skill-specific recommendation resources (which can include, but not limited to technique suggestions, reference textbooks, and other). In addition, by clicking on the graph icon next to the recommendation resources, users can visualize their scenario-specific skill-related proficiency history, as displayed in the right-hand panel of Figure 5. For each skill, performance details are provided (described in Section 4.1 as subskills). Similarly to the global skill-proficiency-recommendation format, for each performance detail, a proficiency level, corresponding recommendation resources and display of performance history are provided.



Rabaska Analysis ?

Recommendations

Skill: Clear blood

Skill: Aspirate tumour

PL: 4	Start with a relatively high aspiration power in the centre of the tumour, to debulk until the healthy underlying cortex is first seen. Refer to: Steinmetz MP, Krishnaney A, Lee JH. Surgical Management of Convexity Meningiomas. In: Neuro-Oncology, Badie, ed. New York: Thieme, 2006:145-152."	

Performance Details

Minimize force on brain	PL: 1	Too much force was applied on the brain. Use the force gauge shown in the overlay to monitor the force being applied on the tissue.	
Minimize force on tumour	PL: 1	Too much force was applied on the tumour. Repeat the following scenarios to improve your instrument handling: Ultrasonic Aspirator 101.	
Minimize removal of brain	PL: 3	Lower the aspiration power either at the source by adjusting the knob. Monitor the aspiration intensity using the gauge shown in the overlay to prevent healthy tissue from being removed.	

Skill: Instrument Handling

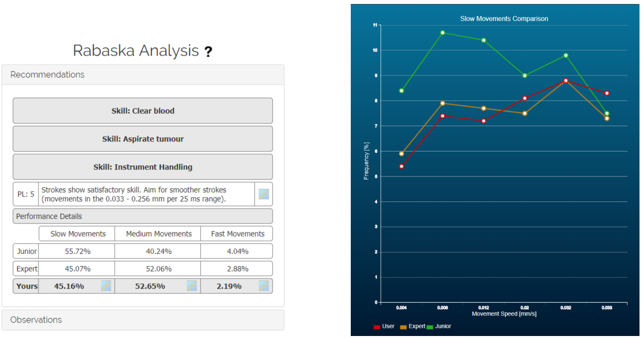
Observations

Proficiency History: Debulking 101, Aspirate tumour, Minimize removal of brain
Proficiency Level
0
1
2
3
4
5
6
7
Attempt
1
2
3
4
5
6
7
8
9
10
11

Figure 5: Recommendations based on the direct analysis module within Rabaska.

Figure 6 displays the integration of the SPM module results, specifically related to instrument movement speed, within the recommendations portion of the *Rabaska Analysis*. The same display format to the direct skill-proficiency-recommendation analysis was used to ensure seamless usability by the users. As such, a global proficiency level was provided, alongside improvement recommendations and an account of proficiency history. Performance details are also provided. In addition, for this advanced analysis, the percentage of movements designated as slow, medium and fast for both junior residents and expert neurosurgeons are presented side-by-side with those of the user for the particular simulation scenario for easy comparison. This constitutes the results extracted from the trained SPM model. By clicking the graph icon for each of the movement speeds, the user can obtain further information regarding the frequency of specific movement speeds in that category, in comparison to both junior and expert users (as shown in the right-hand panel of Figure 6).

Figure 7 displays the integration of the SPM model results, specifically related to the user's common movements, within the observations portion of the *Rabaska Analysis*. The results provide the probability of the user to execute a particular move that is seen (1) predominantly with junior residents and (2) predominantly with expert surgeons. The corresponding probabilities of a junior resident and expert neurosurgeons to execute the same two movements, as extracted from the trained SPM model, is also provided. For each of the common movements described, users can visualize, by clicking on the graph icon, the type of movement, in the corresponding plane, for which the results are provided. For these results, no explicit recommendations are provided, thus justifying their designation as an observation.



Rabaska Analysis ?

Recommendations

Skill: Clear blood

Skill: Aspirate tumour

Skill: Instrument Handling

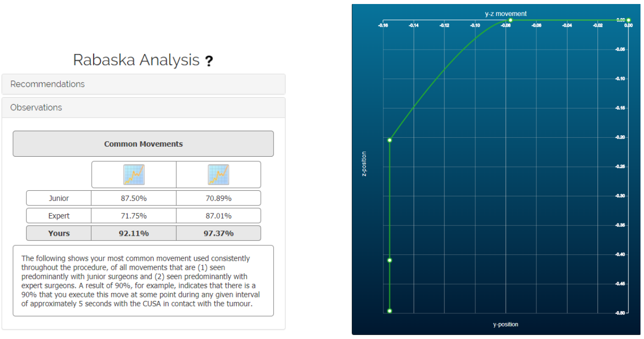
PL: 5	Strokes show satisfactory skill. Aim for smoother strokes (movements in the 0.033 - 0.256 mm per 25 ms range).	📈

Performance Details

	Slow Movements	Medium Movements	Fast Movements
Junior	55.72%	40.24%	4.04%
Expert	45.07%	52.06%	2.88%
Yours	**45.16%** 📈	**52.65%** 📈	**2.19%** 📈

Observations

Figure 6: Recommendations on instrument handling based on the SPM module within Rabaska.



Rabaska Analysis ?

Recommendations

Observations

Common Movements

	📈	📈
Junior	87.50%	70.89%
Expert	71.75%	87.01%
Yours	**92.11%**	**97.37%**

The following shows your most common movement used consistently throughout the procedure, of all movements that are (1) seen predominantly with junior surgeons and (2) seen predominantly with expert surgeons. A result of 90%, for example, indicates that there is a 90% that you execute this move at some point during any given interval of approximately 5 seconds with the CUSA in contact with the tumour.

Figure 7: Common movements feedback to user within the Observation component driven by the SPM module.

5. Discussion and Future Work

As previously discussed, the results of the proposed SPM model (Section 3.2) and consequently those presented in Section 4.4 are in preliminary stages. However it would be worthy to highlight the potential implications that they can bring to the existing training process. The SPM model, with its consequent integration within Rabaska, supports the widely recommended positive encouragement method to learning.

One specific example that demonstrates this notion is through the test trial with a beginner user of NeuroTouch and Rabaska who had no previous simulation experience. Early in the user's training, an overall low score of -9.8/100 was achieved for the Tumour Debulking 101 scenario. However, further investigation into the performance details, provided by the SPM model, revealed that the (junior) user exhibited similar instrument handling behaviour to that of an expert neurosurgeon. After continued practice, the user succeeded in increasing the overall performance score to 60.9/100, which was a significant improvement. However, when consulting the results of the SPM model, the hand movement of the user diverged from what is considered to be a typically expert maneuver. This example demonstrates the potential importance of executing low-level, behaviour-based analysis of users' instrumentation and technique and emphasizes the fact that high-level metrics, and corresponding benchmark values that are proposed under the context of expert information may not be sufficient to assess proficiency.

The fact that no expert information on neurosurgical skill is utilized by the AI engine is a key factor differentiating this work from similar studies. The analysis is entirely uninformed on any sort of baseline behaviour that would be considered representative of skilled or unskilled technique, and relies solely on patterns found in the data to build a model that represents a particular skill class. An obvious caveat to this approach is the reliance on the test subjects to generate representative activity. Thus, frequent behaviours exhibited by experts are considered to be expert behaviours, and so on. However, a critical advantage of the current approach is that there is no need to rely on experts to manually generate or verify behaviours, and thus a much wider variety of behaviours can be identified. To verify the validity of these behaviour patterns truly being indicative of a class of subjects, classification models are built using these patterns, and cross-validation is used to explore whether subjects can be correctly classified using the common behaviours identified.

It is important to note that the interpretation of the results and the corresponding potential impact of the current work are presented with caution and in the context of the limitations. More specifically, the dataset utilized for the SPM model was obtained on a volunteer basis during a large-scale conference. In addition, study participants were asked to self-identify in the context of their skill class. Although several steps were executed in order to pre-process and filter the dataset, the mediocre quality of the dataset is acknowledged. In the context of an AI-driven analysis, the size of the dataset also proves to be preliminary but the presented work was considered as an exploratory pilot study. A large-scale, protocol-driven study would be required, constituting a logical progression of the work, in order to validate the presented results.

While identifying behavioural patterns that are indicative of skill can offer very useful information when incorporated as part of an overall learning and recommendation system, one particular avenue for future work is to facilitate the ability to make the output of the results more personalized and interactive. Rather than statically presenting a number of common movements to the learner that were deemed skillful/unskillful, it would be far more ideal for the learner to be able to investigate certain movements captured in a particular context, and then be able to ascertain what was done well or poorly at that point.

Another opportunity for future work centers on the ability to allow for a subject matter expert or instructor to refine the models constructed by the AI engine. For example, consider a movement that was deemed a junior movement by the AI engine that is in fact a new technique that is being taught only recently, and is thus a very skilled and effective method. The ability to allow feedback from the subject matter expert and facilitate automatic model updates accordingly would even further improve the effectiveness of this sort of analysis as a valuable instructional tool.

One should note also that this type of analysis is not limited to surgical simulators. Simulators from many other domains that are used for training and can capture movements and positional data, such as flight simulators, search and rescue simulators, or even many types of intelligent tutors, could potentially significantly benefit from this sort of low-level behavioural analysis, thus enhancing the learning-teaching dynamic across a wide array of teaching and learning domains.

6. Conclusions

This paper presented the results of an analytics-based study to determine key differences in junior- and expert-level surgical technique when engaging with the neurosurgical simulator NeuroTouch. Frequent pattern-based techniques were used in the study to identify key movement sequences that differentiated surgical skill, and skill classification models were constructed accordingly based on these identified behaviours. The paper makes two key contributions to the field of machine learning in the medical field: 1) rather than build models of skill or behaviour based on high-level summary metrics such as blood loss and average force, a low-level, behaviour-based analysis of surgeons' instrumentation and technique is employed, where frequent patterns in the sequence of movements and changes in positioning are instead identified, and 2) an entirely uninformed analysis that relies solely on patterns found in the data rather than baseline behaviour that would be considered representative of skilled or unskilled technique is conducted such that no expert information on neurosurgical skill is utilized by the AI engine.

References

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

Abdullatif Alwasel, Ali Sabet, Mohammad Nahangi, Carl T Haas, and Eihab Abdel-Rahman. Identifying poses of safe and productive masons using machine learning. *Automation in Construction*, 84:345–355, 2017.

Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435. ACM, 2002.

Hamed Azarnoush, Gmaan Alzhrani, Alexander Winkler-Schwartz, Fahad Alotaibi, Nicholas Gelinas-Phaneuf, Valérie Pazos, Nusrat Choudhury, Jawad Fares, Robert Di-Raddo, and Rolando F. Del Maestro. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *International Journal of Computer Assisted Radiology and Surgery*, 10(5):603–618, May 2015. ISSN 1861-6429. doi: 10.1007/s11548-014-1091-z. URL https://doi.org/10.1007/s11548-014-1091-z.

Hamed Azarnoush, Samaneh Siar, Robin Sawaya, Gmaan Al Zhrani, Alexander Winkler-Schwartz, Fahad Eid Alotaibi, Abdulgadir Bugdadi, Khalid Bajunaid, Ibrahim Marwa, Abdulrahman Jafar Sabbagh, and Rolando F. Del Maestro. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. *Journal of Neurosurgery*, 127(1):171–181, 2017. doi: 10.3171/2016.7.JNS16322. URL https://doi.org/10.3171/2016.7.JNS16322. PMID: 27689458.

Abdulgadir Bugdadi, Robin Sawaya, Duaa Olwi, Gmaan Al-Zhrani, Hamed Azarnoush, Abdulrahman Jafar Sabbagh, Ghusn Alsideiri, Khalid Bajunaid, Fahad E. Alotaibi, Alexander Winkler-Schwartz, and Rolando Del Maestro. Automaticity of force application during simulated brain tumor resection: Testing the fitts and posner model. *Journal of Surgical Education*, 75(1):104 – 115, 2018. ISSN 1931-7204. doi: https://doi.org/10.1016/j.jsurg.2017.06.018. URL http://www.sciencedirect.com/science/article/pii/S1931720417301149.

Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth. Rule discovery from time series. In *KDD*, volume 98, pages 16–22, 1998.

Nicholas Gélinas-Phaneuf, Nusrat Choudhury, Ahmed R. Al-Habib, Anne Cabral, Etienne Nadeau, Vincent Mora, Valerie Pazos, Patricia Debergue, Robert DiRaddo, and Rolando F. Del Maestro. Assessing performance in brain tumor resection using a novel virtual reality simulator. *International Journal of Computer Assisted Radiology and Surgery*, 9(1):1–9, Jan 2014. ISSN 1861-6429. doi: 10.1007/s11548-013-0905-8. URL https://doi.org/10.1007/s11548-013-0905-8.

Antonio Gomariz, Manuel Campos, Roque Marín, and Bart Goethals. Clasp: An efficient algorithm for mining frequent closed sequences. In *Advances in Knowledge Discovery and Data Mining*, pages 50–61. Springer, 2013.

Gregor Kennedy, Ioanna Ioannou, Yun Zhou, James Bailey, and Stephen O'Leary. Mining interactions in immersive learning environments for real-time student feedback. *Australasian Journal of Educational Technology*, 29(2), 2013.

Neal Lesh, Mohammed J Zaki, and Mitsunori Ogihara. Mining features for sequence classification. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 342–346. ACM, 1999.

Neal Lesh, Mohammed J Zaki, and M Oglhara. Scalable feature mining for sequential data. *Intelligent Systems and their Applications, IEEE*, 15(2):48–56, 2000.

Carl H Mooney and John F Roddick. Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR)*, 45(2):19, 2013.

Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 0215–0215. IEEE Computer Society, 2001.

Robin Sawaya, Abdulgadir Bugdadi, Hamed Azarnoush, Alexander Winkler-Schwartz, Fahad E. Alotaibi, Khalid Bajunaid, Gmaan A. AlZhrani, Ghusn Alsideiri, Abdulrahman J. Sabbagh, and Rolando F. Del Maestro. Virtual reality tumor resection: The force pyramid approach. *Operative Neurosurgery*, page opx189, 2017. doi: 10.1093/ons/opx189. URL http://dx.doi.org/10.1093/ons/opx189.

Christopher Sewell, Dan Morris, Nikolas H Blevins, Sanjeev Dutta, Sumit Agrawal, Federico Barbagli, and Kenneth Salisbury. Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery*, 13(2):63–81, 2008.

Ramakrishnan Srikant and Rakesh Agrawal. *Mining sequential patterns: Generalizations and performance improvements*. Springer, 1996.

Ben-Manson Toussaint and Vanda Luengo. Mining surgery phase-related sequential rules from vertebroplasty simulations traces. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 35–46. Springer, 2015.

Ke Wang and Jye Tan. Incremental discovery of sequential patterns. In *1996 ACM SIGMOD Data Mining Workshop: Research Issues on Data Mining and Knowledge Discovery (SIGMOD96)*, pages 95–102, 1996.

Alexander Winkler-Schwartz, Khalid Bajunaid, Muhammad A.S. Mullah, Ibrahim Marwa, Fahad E. Alotaibi, Jawad Fares, Marta Baggiani, Hamed Azarnoush, Gmaan Al Zharni, Sommer Christie, Abdulrahman J. Sabbagh, Penny Werthner, and Rolando F. Del Maestro. Bimanual psychomotor performance in neurosurgical resident applicants assessed using neurotouch, a virtual reality simulator. *Journal of Surgical Education*, 73(6):942 – 953, 2016. ISSN 1931-7204. doi: https://doi.org/10.1016/j.jsurg.2016.04.013. URL http://www.sciencedirect.com/science/article/pii/S1931720416300265.

Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010.

Junfu Yin, Zhigang Zheng, and Longbing Cao. Uspan: an efficient algorithm for mining high utility sequential patterns. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 660–668. ACM, 2012.

Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2):31–60, 2001.