Transformer-based intelligent tutoring system for surgical training optimization

Abicumaran Uthamacumaran, BSc

Department of Surgical and Interventional Sciences

Faculty of Medicine and Health Sciences

McGill University

Montreal, Canada

June 2026

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

Supervised by Dr. Rolando F. Del Maestro

TABLE OF CONTENTS

ABSTRACT

Translation of intelligent tutoring systems to operative settings remains challenging due to the sequential nature of procedures and data limitation. The deep learning architecture which best optimizes improvement of these educational systems for surgical training remains unknown. This investigation therefore evaluated Long Short-Term Memory (LSTM) networks, Transformers, and hybrid architectures to develop individual calf brain Ex Vivo Intelligent Continuous Expertise Monitoring Systems (EV-ICEMS). This case series study included 47 participants, novices, junior and senior residents/fellows along with neurosurgeons who performed three subpial ex vivo corticectomies on the calf brain model. Primary outcomes included composite expertise scores (–1.00 [novice] to 1.00 [expert]) and assessment of each model's construct and predictive validity. The Transformer and LSTM-Transformer demonstrated better construct validity than other architectures but only the Transformer differentiated seniors from juniors (mean difference, 0.59; 95% CI, 0.23 to 0.96; P = .004). The Transformer demonstrated superior predictive validity (R^2 = 0.424; MSE = 0.168; MAE = 0.365) with the steepest score-year slope (0.150; 95% CI, 0.074 to 0.225). This study outlines a deep learning methodology to differentiate surgical expertise levels and optimize the development of intelligent tutoring systems for surgical training in realistic operative environments.

RÉSUMÉ

La traduction des systèmes de tutorat intelligent vers des environnements opératoires reste difficile en raison de la nature séquentielle des procédures et des limitations des données. L'architecture d'apprentissage profond qui optimise le mieux ces systèmes éducatifs pour la formation chirurgicale reste inconnue. Cette investigation a donc évalué les réseaux à mémoire à

long et court terme (LSTM), les Transformeurs, et les architectures hybrides afin de développer des Systèmes Individuels de Surveillance Continue de l'Expertise Ex Vivo sur cerveau de veau (EV-ICEMS). Cette étude de série de cas a inclus 47 participants, soit des novices, des résidents/fellows juniors et seniors ainsi que des neurochirurgiens qui ont effectué trois corticectomies sous-piales ex vivo sur un modèle de cerveau de veau. Les critères de jugement principaux comprenaient les scores composites d'expertise (–1,00 [novice] à 1,00 [expert]) et l'évaluation de la validité de construit et la validité prédictive de chaque modèle. Le Transformeur et le LSTM-Transformeur ont démontré une meilleure validité de construit que les autres architectures, mais seul le Transformeur différenciait les seniors des juniors (différence moyenne, 0,59 ; IC à 95 %, 0,23 à 0,96 ; P = 0,004). Le Transformeur a démontré une validité prédictive supérieure (R^2 = 0,424 ; MSE = 0,168 ; MAE = 0,365) avec la pente score-années d'expérience la plus prononcée (0,150 ; IC à 95 %, 0,074 à 0,225). Cette étude décrit une méthodologie d'apprentissage profond pour différencier les niveaux d'expertise chirurgicale et optimiser le développement de systèmes de tutorat intelligent pour la formation chirurgicale dans des environnements opératoires réalistes.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. Rolando F. Del Maestro, for being an excellent mentor and supervisor who always goes above and beyond. I thank him for allowing me to join the Neurosurgical Simulation and Artificial Intelligence Learning Centre and for the unwavering support, care, and rigour that have shaped this work and my growth. I am deeply grateful to be under his wing and contribute to his vision.

I am grateful to the members of my research advisory committee, Dr. Amir Hooshiar, Dr. Denis Laroche and my chair, Dr. Sampath Kumar Loganathan, for their guidance, feedback, and generosity with their time.

To my collaborators at the Neurosurgical Simulation and Artificial Intelligence Learning Centre — Bianca Giglio, Dr. Matheus Ballestero, Dr. Abdulrahman Almansouri, Dr. Recai Yilmaz, Nour Abou Hamdan, Dr. Houssem-Eddine Gueziri, Dr. D. Louis Collins, Dr. José A. Correa, and all others — thank you for the countless hours of discussion, data collection, and critical feedback that made this thesis possible. A special thanks to Bianca and Dr. Ballestero for their unconditional support in helping me succeed.

I also wish to thank the medical students, neurosurgical residents, fellows, and attending neurosurgeons who volunteered their time to participate in this study.

I thank my parents and siblings for their unconditional love, sacrifices, and support throughout my entire life. This thesis is dedicated to my mother Kalaimathy Arumugam and my father Sivasubramaniam Uthamacumaran.

Finally, I wish to thank all the children in B7 Oncology Wing at the Montreal Children's Hospital who gave me meaning and purpose.

AUTHOR CONTRIBUTIONS

The candidate led the study presented in this thesis and contributed to all aspects of the AI model-derived work, including AI model conceptualization, study design, participant recruitment, deep learning model development and training, statistical analysis performed for the model comparisons under statistical guidance, results interpretation, and manuscript writing.

Dr. Abdulrahman Almansouri and Nour Abou Hamdan contributed to the development of the ex vivo calf brain subpial corticectomy simulation model and carried out the trial and acquired all the data on which my contribution is based. The candidate had full access to all instrument tracking data previously acquired by the team and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Dr. Recai Yilmaz contributed to study conceptualization, deep learning methodology, and results interpretation.

Bianca Giglio and Dr. Matheus Ballestero contributed to manuscript writing, drafting and critical revision.

Dr. Houssem-Eddine Gueziri contributed to the development related to the ex vivo model.

Dr. D. Louis Collins contributed by allowing the trial to be carried out in the experimental Operating room in his lab along with expertise on running the trial

Dr. José A. Correa provided statistical methodology, statistical analysis guidance, and critical review of the statistical reporting.

Dr. Rolando F. Del Maestro conceived and orchestrated the project, supervised the project and provided the neurosurgical expertise, ethics approval, and resources necessary to write, design, conduct, and interpret the study.

ABBREVIATIONS

3D — Three-Dimensional

AERA — American Educational Research Association

AI — Artificial Intelligence

ANOVA — Analysis of Variance

APA — American Psychological Association

AR — Augmented Reality

CBME — Competency-Based Medical Education

CI — Confidence Interval

CONSORT-AI — Consolidated Standards of Reporting Trials — Artificial Intelligence

EPA — Entrustable Professional Activity

EV-ICEMS — Ex Vivo Intelligent Continuous Expertise Monitoring System

ICEMS — Intelligent Continuous Expertise Monitoring System

LSTM — Long Short-Term Memory

MAE — Mean Absolute Error

ML — Machine Learning

MSE — Mean Squared Error

NCME — National Council on Measurement in Education

OR — Operating Room

OSATS — Objective Structured Assessment of Technical Skills

PGY — Post-Graduate Year

PLA — Polylactic Acid

RCT — Randomized Controlled Trial

RNN — Recurrent Neural Network

SD — Standard Deviation

VOA — Virtual Operative Assistant

VR — Virtual Reality

THESIS INTRODUCTION

In surgery, technical skill performance is a critical determinant of patient safety, recovery, and quality of life.[1-5] This complex relationship is especially unforgiving in neurosurgery, wherein small deviations in instrument trajectory, coordination, or tissue handling can lead to lifelong health consequences, suffering and mortality.[2,3,5] For this reason, technical skill acquisition and evaluation among surgical trainees is a central safety concern with the purpose of surgical education to mitigate surgical errors.[1-5]

Traditional surgical training has long relied on supervised participation in patient care, with accountability increasing as trainees demonstrate clinical readiness.[6-10] Surgical curricula remain poorly standardized across programs, with differential exposure to surgical procedures determined by case availability, and the operating room (OR) is a stressful and challenging educational environment for novice learners, limiting deliberate practice.[11-15] Competency-based medical education (CBME) frameworks and assessment tools such as the Objective Structured Assessment of Technical Skills (OSATS) and entrustable professional activities (EPAs) are widely used standards for expertise evaluation.[14-24] Thus, much technical assessment still relies on episodic, qualitative, and subjective observations vulnerable to inter-rater variability rather than continuous quantitative evidence.[14-24]

Parallel advances in virtual reality (VR)-powered simulation and surgical AI have created a robust quantitative assessment paradigm in which operative performance can be recorded and analyzed as high-resolution behavioral data.[25-32] Prior work from our group established that intelligent tutoring systems, deep-learning-driven pedagogical platforms that continuously assess operative performance and deliver real-time, coachable verbal feedback with performance metric-derived scoring, can improve bimanual psychomotor skills learning in VR outperforming

teaching by human expert instructors.[25-30] The Intelligent Continuous Expertise Monitoring System (ICEMS), a Long Short-Term Memory (LSTM)-based deep learning system that assesses surgical performance at 0.2-second intervals and delivers continuous expertise scores with personalized AI-guided feedback, is one such platform.[28,29] Through a series of randomized controlled trials (RCTs) and crossover studies, ICEMS has provided strong evidence that AI-derived performance data can improve surgical simulation training, when integrated with expert instruction.[27-29]

The next translational step is to develop, test and validate the ICEMS into more biologically realistic operative conditions.[33-35] VR remains constrained by simulated physics and therefore cannot fully replicate the haptic, kinematic, and instrument-handling demands of real operative tissue environments.[33-35] Ex vivo calf brain models developed by our group provide a deliberately intermediate platform that preserves biological tissue fidelity with anatomical realism, permits use of real operative instruments (microscissors, bipolar forceps, and ultrasonic aspirator), and supports continuous fiducial-based instrument tracking, while retaining a safer, reproducible, and more standardized than patient-based training.[33-35] These platforms are positioned to develop a bridge between VR simulation and the eventual development of an "intelligent OR" capable of continuous intraoperative assessment, real-time monitoring as early warning prediction systems, and error mitigation.[33-35]

The major question in this area concerns model architecture.[36-42] The original VR-based ICEMS employed an LSTM network, which is well suited to modelling local and sequential dependencies in time-series performance data.[38,42] However, it may not optimally capture the long-range, global relationships between multiple instruments acting in coordination during a complex operative procedure.[36,41,42] Transformer architectures now form the foundation of many

state-of-the-art AI systems, including large language models and chatbots, use attention mechanisms to evaluate relationships across sequences in parallel. These mechanisms explicitly model long-range temporal dependencies and higher-order interdependencies emerging from complex interactions.[36,37] Transformers underpin contemporary foundation models, that is, large models pre-trained on massive datasets and subsequently adapted to specific tasks. For instance, SurgVISTA is a self-supervised surgical video foundation model pre-trained on millions of surgical frames to learn transferable representations, which were then adapted and evaluated across diverse surgical tasks, including phase recognition, workflow understanding, action recognition, and surgical skill assessment.[43] Transformers have also begun to demonstrate advantages over LSTM-based architectures in medical AI tasks where contextual information, data scale, and global structure are critical.[37-40,43] However, it remains uncertain whether recurrent LSTM architectures, which have shown strong performance in VR-based intelligent tutoring systems, remain optimal for modeling the greater complexity and emergent behavioral dynamics of multi-instrument coordination in performance data from a realistic ex vivo setting.[34,36-42]

We therefore hypothesized that a Transformer, with its capacity for parallel processing and explicit modelling of long-range temporal relationships, may confer architectural advantages in this context. Whether a Transformer, an LSTM, or a hybrid combination of the two best supports an intelligent tutoring system trained on an ex vivo calf brain model has never been tested.[34,36-42]

This thesis addresses these knowledge problems.[34,36-42] Using kinematic data from an ex vivo calf brain subpial corticectomy model, the work develops Ex vivo-ICEMS (EV-ICEMS) and compares LSTM, Transformer, and hybrid architectures under a common evaluation framework.[36-42] To empirically test this hypothesis, this study evaluates, for each model, whether

model-derived expertise composite scores separate training groups (i.e., construct validity) and generalize to unseen participants across training years (i.e., predictive validity) to support their use as objective performance measures.[44-47] These findings help inform how intelligent tutoring systems can be built for realistic surgical training environments. They also support the long-term goal of developing an intelligent operating room that can continuously assess surgical skill, monitor performance, and help mitigate errors during procedures.[48-52]

BACKGROUND

From Apprenticeship to Quantitative Surgical Training

Modern surgical training traces its origin to the residency program established by William Stewart Halsted at Johns Hopkins Hospital in 1890.[6-8] His paradigm advanced surgical education from passive observation to supervised practice to instruction, catalyzing widespread global adoption, and it remains the structural backbone of contemporary surgical education.[6-10]

The limitations of this pedagogical framework are now well characterized.[11-15,53] As operative exposure depends on the clinical availability of patients, surgical curricula and training experiences vary substantially across residency programs.[11–13,53] This limits clinical preparedness for real-life complexity, risks, and uncertainty. The OR compounds this problem as its primary obligation is patient-centered care and safety, not deliberate practice, teaching, or risk-free failure, wherein most resilience and learning develop. Reduced operative time, duty-hour restrictions, and expanding team-based responsibilities further compress opportunity windows for hands-on procedural skill acquisition.[11-13,53]

Neurosurgery represents one of the most technically demanding and highest-risk specialties in surgical education.[2,14,16,17] The brain, the most advanced complex system known to humankind, is a multilayered network of fragile and sensitive structures, spanning vascular architecture to the intricate highways of the neural connectome, creating multiscale risks for surgical error.[2,16,17] Prospective data report substantial adversity rates, with nearly one-fifth emerging from procedural complications, underscoring the vulnerability of this patient population.[2,3] Such events include neurological deficits, infections, hemorrhages, and the need for revision surgery, each with long-term consequences.[2,3,14] Accordingly, neurosurgical technical performance is a critical determinant of patient outcomes, influencing perioperative mortality, long-term disability, quality of life, and healthcare system burden.[2,5,16,17] Even minor technical deviations, such as a millimetre of resection margin, improper tissue handling or vessel preservation, can result in poorer prognosis or lasting functional impairment.[2,5,16,17] For instance, in tumor resection, epilepsy surgery, and neurovascular interventions, timely and precise surgery improves quality of life and long-term survival and, when combined with adjuvant therapies, reduces recurrence, disease progression, and adversities such as rebleeding or stroke.[2,3,5,54,55]

CBME emerged as a solution framework in response to the variability of apprenticeship-based training.[14-19] In Canada, the Royal College of Physicians and Surgeons of Canada implemented this transition through EPAs, which specify the clinical goals or tasks trainees must demonstrate competence in prior to surgical practice.[14-19] However, CBME remains entwined with some of the very limitations it emerged to address, including the use of tools such as OSATS and narrative assessments, which can remain qualitative, subjective, and prone to systemic biases or power inequities.[17,20-24] The unmet need and what this thesis advocates for, are objective, data-

driven precision systems that complement and extend CBME capable of continuous, reproducible and standardized technical-skill quantification in surgical training.[20-24]

Simulation Training in Surgical Education

In surgical education, simulation offers what the OR cannot: the ability to make and correct errors without harming patients, refine procedural skills, and adaptive learning against well-defined objectives without the consequences of real-world mistakes.[56-60] Simulation platforms span across biologically realistic systems (e.g., cadaveric, in vivo animal models, ex vivo animal tissue) and synthetic platforms (e.g., 3D-printed, VR, and augmented reality [AR]).[56-61]

Synthetic simulators are affordable, accessible, and portable risk-free settings allowing repetition, but generally trade-off realism and are constrained in procedural complexity.[58-63] VR simulators are at the heart of contemporary surgical education, combining multi-sensory (visual, auditory, and, increasingly, haptic) feedback to immerse the trainee with autonomous capture of large quantities of performance data.[59-64]

VR's core advantage in surgical education is data-driven: every aspect of a trainee's performance, such as instrument trajectory, kinematic and force profiles, volume of blood loss, extent of tissue injury, and other safety-oriented metrics, is recorded as a quantitative time-series for computational analyses.[57-59,64] Simulation-based training is grounded in deliberate practice, involving systematic, repetitive performance with personalized feedback, and in experiential learning theory.[15,60-65] Contemporary work expands this to include embodied affective and cognitive dimensions, as emotional regulation and cognitive load show how instruction, task design, and the learning environment influence performance.[65,66] From a biopsychosocial-cultural model of medicine, the trainee, educator, and learning environment can be viewed as

interacting components of a complex adaptive ecosystem.[70] Tools such as the Medical Emotions Scale (MES) and Cognitive Load Index (CLI) have been incorporated into recent simulation RCTs to characterize these multiscale dimensions as causal co-predictors of technical performance.[65-69,71]

A prospective blinded controlled trial demonstrated that laparoscopic skills acquired on a VR simulator transferred to improved performance on a real-world laparoscopic cholecystectomy, supporting the broader body of evidence that VR-based skill acquisition can translate to real operative performance.[72] This transfer of skill has been extended to neurosurgery, where simulator-trained residents demonstrated improved performance in live ventriculostomy procedures and endoscopic endonasal surgery compared to untrained controls.[73,74] Our group's RCT demonstrates that VR-based surgical training not only improves simulated performance but transfers to summative and more complex tasks, with AI-augmented personalized instruction achieving significantly higher expertise scores in the realistic resection scenario. [26-30]

Despite their strengths, VR simulators have an irreducible limitation: they do not use the actual operative instruments and cannot fully replicate the real-world procedural complexity and tactile properties of biological tissue.[33-35] This limitation is the direct motivation for ex vivo models, which occupy the transitional bridge between VR and the actual OR environment.[33-35,75]

Neurosurgical Simulation Training

Neurosurgical education may benefit from simulation training as procedures are technically complex, the cost of error is high, and supervised practice or operative exposure remains limited.[76-80] Our group has contributed to the development and validation of multiple neurosurgical simulation platforms, including NeuroVR (CAE Healthcare, Montreal, Canada)

and ex vivo calf brain models.[26-29,33–35,64,75,81,82] The NeuroVR is a high-fidelity VR simulator that serves as a major data-driven platform for this research program. [26-29,64] It enables continuous performance-data capture from simulated scenarios, including the subpial tumor resection task central to the RCTs validating ICEMS and its predecessor, the Virtual Operative Assistant (VOA). [25–30,64,81,82] The platform recreates the microsurgical environment through stereoscopic binocular microscope, with bipolar forceps and an ultrasonic aspirator mounted on haptic interfaces for improving bimanual psychomotor skill acquisition and training. [25,28,64,81,82] Instrument movements and surgical behaviors are captured at high temporal resolution and continuously monitored. [25,28,64,81,82] The platform has demonstrated face, content, and construct validity as a surgical training tool. [25,28,64,76-82] Face validity assesses whether the model appears realistic to users; content validity refers to whether experts judge the task and metrics to be representative of the simulated surgical skill; construct validity refers to whether the model's ICEMS scores differentiate expertise groups; and predictive validity refers to whether model-derived scores generalize to previously unseen performance data across training years. [28,34,45–47]

However, translating intelligent tutoring systems beyond the NeuroVR requires a risk-free model that uses real operative instruments on biological tissue while upholding patient safety. This motivated the development of an ex vivo calf brain corticectomy model and, in parallel, steers the present thesis — whether AI can teach technical skills in a realistic setting that more closely resembles the human OR.[33,34]

The Subpial Resection Technique: A Model Task for Bimanual Skill Assessment

The subpial resection is a bimanual neurosurgical technique for the removal of pathological cortical tissue central to brain tumors and epileptogenic foci, while preserving the pia mater and minimizing injury to surrounding healthy tissue.[78-81] The technique was initially described by

Penfield and Jasper in the context of epilepsy surgery and has since become a standard procedural intervention in both epilepsy and tumor neurosurgery.[83-85] The extent of resection is a predictor of patient survival. For instance, in high-grade glioma, greater extent of supratotal resection, has been associated with improved survival. [54,55,85,86]

The procedure involves the coordinated use of three instruments.[34,80,81] Microscissors are used to make an initial incision of the pia.[34,80,81] Bipolar forceps are used to gently elevate and hold the pia from the underlying cortex and, in the operative setting, to cauterize small bleeding vessels.[34,80,81] A suction device or ultrasonic aspirator is used to resect the underlying pathological tissue while preserving the elevated pia and surrounding structures, including adjacent gyri, blood vessels, and the subcortical white matter at the depth of the resection cavity.[34,80,81] The targeted depth of white matter entry is pathology-dependent: epilepsy procedures may follow a subpial plane around an epileptogenic focus, whereas tumor surgery may require deeper or wider resection margins as determined by tumor infiltration, vasculature, functional boundaries, and patient safety.[55,84,85] In our present ex vivo model, the task was standardized to three superficial subpial resection targets without the patient-level complexity. Therefore, EV-ICEMS assessed instrument-motion patterns throughout the task, but did not independently score white-matter depth, residual lesion, vascular injury, or tissue-safety outcomes. In the OR and in VR simulators, the technique additionally involves careful control and avoidance of vascular injury or bleeding. In ex vivo non-perfused models, bleeding is not simulated, but the bimanual kinematic demands of pial elevation and subpial aspiration, as well as the precisely timed and coordinated movements of the scissors and bipolar are preserved.[28,34,85,86]

Thus, subpial resection is an appropriate and reproducible learning task for EV-ICEMS development because its core behaviors can be measured as teachable, motion-based kinematic metrics: velocity, acceleration, jerk, and inter-instrument distance, that can be extracted from instrument-tracking data and used as training input to the deep learning model.[34] Further, it embodies a spectrum of emergent behavioral data such as tissue handling and dexterity, bimanual coordination, regulation of force surrogates (as measured by the kinematics), and precise spatiotemporal control of instruments with respect to delicate structures.[20,28,34,82]

Ex Vivo Calf Brain Models in Neurosurgical Education

Ex vivo animal brain models occupy the transitional space between VR simulators and live-patient OR.[33,34] They allow the interaction of actual operative instruments with tissue biophysics, preserving haptic feedback, anatomical realism, and mechano-tactile resistance.[33,34,87,88] Further, they enable reproducibility, standardization, and the absence of patient risk unlike live-patient surgery.[33,34,87,88] Our group has also published best-practice guidelines for the use of ex vivo brain models in neurosurgical education.[33,34,87,88]

The calf brain was chosen for our subpial resection model because it is morphologically similar to the human pediatric brain, widely available, inexpensive, and anatomically suitable for training microneurosurgical techniques. Almansouri and colleagues developed a calf brain corticectomy model in which a fresh calf brain is positioned within a human skull replica with an off-midline craniotomy and operated on under an OPMI Pico surgical microscope.[34] The instruments are equipped with 3D-printed polylactic acid (PLA) mounts carrying fiducial markers, enabling continuous real-time 3D tracking of each instrument via a dual-camera infrared optical system (FusionTrack 500 / Polaris). Tracking data are streamed through the PLUS/OpenIGTLink pipeline into 3D Slicer for acquisition and analysis. [34,75,89]

This platform has demonstrated face and content validity in a case-series validation study involving skilled neurosurgeons and fellows and less-skilled senior residents and fellows, with median Likert scores of 6.0/7 and no significant differences between skilled and less-skilled groups.[34] The model was rated as highly appropriate for training the subpial resection technique, and the authors recommended its integration into surgical curricula.[34] The next step in validation would be to establish the model's construct and predictive validity, the core objectives of the study presented in this thesis.[45-47]

Ex vivo models have additional strategic importance for intelligent tutoring systems development.[33-35] VR-based intelligent tutoring systems such as ICEMS were trained on rich performance data including simulated bleeding, simulated tissue injury, and force surrogates derived from haptic handles — variables that are not available in a non-perfused ex vivo model.[28,29,34,64] The transition from VR to ex vivo thus requires the adoption of kinematic performance metrics (velocity, acceleration, jerk, inter-instrument distance) as the primary features on which an AI tutoring system must learn to discriminate expertise. Kinematic metrics are transparent, universal across instruments and platforms, and do not depend on simulator-specific physics engines. This common ground allows an AI model accurately trained on ex vivo kinematic data to be a prerequisite and template for an intelligent tutoring system that will eventually function in the human OR.[20,28,33-35,44,90-93]

Validation Framework for EV-ICEMS

A validation study asks whether a tool or system is fit for a particular purpose with a target population.[45-47] For EV-ICEMS, validity concerns whether the AI model-derived scores can distinguish surgical expertise-based kinematic and behavioral patterns. Two validity frameworks are in current use: a traditional framework, which entails discrete "types" of validity, and a

unified contemporary framework in which validity is understood as a single argument from multiple sources of evidence (Table 1).[28,34,45-47,94,95]

The traditional framework organizes validity into various categories, namely face, content, construct, and predictive validity components. Messick's contemporary framework reconceptualizes validity as an irreducible evidence-based framework. A model or tool's use for a given purpose is valid to the extent that evidence supports that use.

This thesis adopts the traditional framework to describe the validity of the EV-ICEMS to maintain continuity with our group's prior ICEMS and simulation-validation studies. Specifically, we assess its construct validity (the ability of a given deep learning architecture to differentiate between the groups: novices, juniors, seniors, and experts, based on composite expertise scores) and its predictive validity (the ability of the model to generalize to previously unseen data). In this context, construct validity assigns value to the intended purpose and use of the AI model while the predictive validity denotes whether the model-derived composite scores can predict juniors and seniors across their years of training. This validity assessment allows closure with prior work on the face and content validity of the ex vivo calf brain model and direct comparison on the construct and predictive validity of the original VR-based ICEMS. The findings of this thesis arguably also supply several sources of evidence within Messick's contemporary lens: the input kinematic metrics represent test content grounded in expert practice; the comparison of composite scores across training years supplies evidence of relationship to other variables; and the objective of optimizing an intelligent tutoring system for error mitigation encompasses consequences.[45-47,94,95]

Performance Assessment Methods in Surgical Education

Assessment of surgical technical skill is central to competency-based training.[17,20-24] OSATS remains the primary rating tool, scoring trainee technical performance across seven distinct behavioral domains using a five-point Likert scale.[17,20,21] However, as discussed, OSATS and similar measures of expertise like EPAs rely on qualitative human subjectivity.[14,17-24] They do not quantify safety, complex inter-instrumental dynamics, or the fine kinematic modulation of bimanual coordination as pertaining to EV-ICEMS development. Inter-rater variability and evaluator bias are well documented, and there are institutional and regional differences in surgical practice questioning OSATS' robustness.[17,20-23] EPAs benchmark what must be demonstrated but do not resolve these assessment-reliability limitations.[14,18,19,24] This necessitates quantitative, evidence-based systems to complement subjective-based instruments like OSATS.[20,28,31,32,96-100]

The emergence of AI has evolved how surgical competence is assessed, moving away from OSATS toward decentralized, continuous, and quantitative frameworks.[31,32,96-99] Supervised machine learning (ML), a subset of AI models for data-driven pattern recognition, learning, and prediction, trains on labelled examples of novice and expert performance to predict the expertise from previously unseen (unlabelled) performance data.[31,32,96-99]

The standard ML approach is feature-analysis, or metric extraction: a task-specific set of performance metrics is derived from the tracking data, and a supervised ML model (such as an artificial neural network) is trained to classify or regress expertise from these metrics. [31,32,96-100] Metric-based assessment has two main advantages. First, it is granular: a VR simulator can score performance at high temporal resolution throughout the trial, producing a continuous performance trajectory rather than a single end-of-trial score. [25-29,31,32,64,82] Second, it is

explainable: because the model output is derived from interpretable performance metrics, educators and trainees can identify why a score was assigned and which specific behaviors should be modified.[25,28,31,32] Explainability and hence, transparency, is a prerequisite for trust and safety in surgical AI-based assessment and education, where "black box" AI systems are ethically and practically untenable.[25,31,32,52,91,92] The ICEMS described in the next section is the culmination of this line of work.[27-29,31,32] This thesis addresses this problem by restricting EV-ICEMS inputs to ten clinically interpretable, coachable kinematic metrics; validating model outputs against expertise level and training year; and reporting tool-metric analyses, and feature importance analyses, including attention-based salience and integrated-gradient attribution maps, (see Appendix C) to show which instrument behaviors contributed to each model's predictions at the level of the neural network architectures rather than solely relying on the model's machine learning performance metrics or predictive accuracy. [25,28,31,32,36,52,91,92] These explainability analyses quantify each input metric's contribution to the model output and highlight what the model learned within the neural-network representation, not merely how well it predicted (Appendix C). Hence, the AI model is not presented as a "black-box" prediction machine, but rather as an interpretable model whose output scores are read alongside the educationally meaningful specific tool-metric patterns and group-level validity.

Intelligent Tutoring Systems

Intelligent tutoring systems are data-driven AI platforms employed to assess learners and deliver targeted, real-time verbal feedback to mitigate surgical errors and improve technical skill acquisition.[100-107] The first major AI tutor system in neurosurgical training was the VOA developed by our group, which used NeuroVR-derived safety metrics to quantify performance and provide post-trial feedback.[25] An RCT from our group evaluating the VOA demonstrated that

trainees receiving post-hoc AI-metric feedback outperformed those taught by human expert instructors.[26] Cohort data also revealed that embedding intelligent tutoring into a simulation-based surgical curriculum produced unintended consequences in efficiency-related performance domains, suggesting that AI instruction should not be implemented without the mentorship of human surgical educators.[54]

The VOA provides only post-trial feedback. The Intelligent Continuous Expertise Monitoring System (ICEMS) was developed to address this limitation by operating during the task as a real-time coaching and early warning system.[27-29] ICEMS is an LSTM-powered system that continuously assesses bimanual surgical performance at 5 hertz sampling rate (every 0.2 second intervals) in the form of a composite expertise score on a –1 (novice) to 1 (expert) scale.[28] The composite score is computed from aggregating five monitored metrics (healthy tissue injury risk, bleeding risk, instrument tip separation distance, bipolar force, ultrasonic aspirator force).[27-29] The AI system flags a performance deviation whenever a monitored metric exceeds a one-standard-deviation relative to benchmarked expert reference for a continuous period, at which the ICEMS delivers brief verbal feedback to correct the behavior and prevent future errors.[27-29] ICEMS has been validated in VR against surgical expertise levels ranging from medical students to attending neurosurgeons and has been shown to accurately track skill acquisition throughout a neurosurgical training program.[27-29]

Our group's earlier RCT showed that trainees tutored by ICEMS outperformed those tutored by a human expert who was blinded to the ICEMS error data.[29] A randomized cross-over trial subsequently showed that the ordering of instructional modalities, namely of AI and expert instruction, is of critical importance in shaping learning outcomes.[30] Sustained performance improvements emerged only when human expert teaching preceded AI tutoring, whereas

initiating training with the AI tutor followed by expert instruction led to performance deterioration.[30] Most recently, Giglio and colleagues conducted a three-arm RCT in which the human expert instructor delivered personalized, AI-augmented feedback using the ICEMS error data.[27] This AI-enhanced personalized instruction resulted in superior surgical performance and superior skill transfer compared with either ICEMS alone or expert instruction using ICEMS-identical verbal feedback.[27] A complementary investigation by Davidovic and colleagues demonstrated that AI-augmented human instruction alters not only performance but also the frequency and structure of feedback during training.[105] A systematic review with quantitative synthesis of automated feedback systems in surgical training has further affirmed the educational impact of these technologies.[108]

Across this body of work, two themes recur. First, intelligent tutoring systems produce measurable and incremental improvement in trainee learning comparable to, and often exceeding, those produced by human expert instruction alone in VR-based surgical training.[25–30,54,108] Second, the maximal learning outcomes are obtained when the quantitative performance data from intelligent tutoring systems are paired with personalized instruction from human educators.[27,30,105] Both conclusions frame the translational problem at the heart of this thesis: can an intelligent tutoring system be built for a realistic, ex vivo surgical model, and which deep learning architecture best supports it? Addressing these knowledge gaps will steer the next-generation of surgical AI-powered curricula.[27,30,105,108-113]

Deep Learning Rationale: LSTM, Transformers, and Hybrid Models

In complex, realistic ex vivo surgical settings, sequential time-series models may be limited in their ability to capture global patterns, long-range dependencies, and higher-order relationships embedded in behavioral performance data during simulation training. As a result, emergent,

relational coordination patterns, complex dynamics, and holistic representations of surgical expertise may not be fully encoded by the kinematic metrics in isolation.[34,36-40]

Two branches of deep learning architectures currently dominate temporal modelling of complex dynamical systems in data science and medicine: Long Short-Term Memory (LSTM) recurrent neural networks and Transformers. LSTM networks, introduced by Hochreiter and Schmidhuber in 1997, are a class of recurrent neural networks, deep learning algorithms designed to forecast temporal sequences by means of gated memory cells.[38] A gated memory cell is a neural network unit that decides what information to keep (input gate), what to discard (forget gate), and what to pass forward (output gate) to the next time point state in a time-series sequence.[28,38] For instance, during a subpial corticectomy sequence, the LSTM may retain a consistent pattern of smooth bipolar movements, filter out spontaneous spikes in positional/tracking noise or abrupt instrumental jerk, and pass forward only the relevant motion memory needed to predict whether the subsequent motion of that tool-metric resembles expert or novice performance.[28,38]

LSTMs process input sequentially: at each time step they update an internal state that carries information arbitrarily far back in the sequence and in the process may also forget non-salient features or weights.[28,38] LSTMs excel well when training data are limited and sequences are short.[28,38] The original VR-based ICEMS was built on an LSTM.[27-29] Its input layer is a stream of 16 performance metrics sampled at 0.2-second intervals, fed through LSTM layers that capture the temporal evolution of performance, a fully connected layer, and a regression output layer that generates a continuous composite expertise score.[28] A neural network layer is an information processing module or stage composed of mathematical units (neurons) that apply weighted transformations to their input data. The input layer receives the numerical kinematic performance metrics, the LSTM layer sequentially learns time-dependent movement patterns across

successive kinematic states, the fully connected layer combines the learned signal representations across all tool-metrics, and the output layer converts the learned representations to one continuous expertise score.[28,38]

In contrast, Transformers, introduced by Vaswani et al. in 2017, represent a fundamentally newer approach to predictive modelling in time-series data.[36] Rather than processing the data step-by-step, a Transformer employs a self-attention mechanism that computes, in parallel, the pairwise statistical relevance of every element in the time-series to every other element.[36] This holistic representation learning allows the model to adaptively learn long-range, global interdependencies and hence, the latent complex dynamics of the time-series — for example, how an instrument's behavior in the first minute of a procedure relates to its behavior in the last minute — without the signal having to propagate through many recurrent steps like an LSTM network.[36,38] The attention mechanism also allows the model to integrate information across multiple parallel input streams as a multimodal AI system: in a multi-instrument operative procedure, attention can map dynamics on one instrument with events on another irrespective of their relative timing.[36] Self-attention allows the Transformer model to identify which earlier or later events within the same participant trial are most informative, or salient, for the current prediction. For example, an aspirator acceleration pattern later in a trial can be interpreted in relation to earlier microscissor incision behavior or bipolar pial elevation, allowing the model to learn expertise-relevant emergent patterns across instrument metrics.

Transformer-based architectures currently underpin many foundation models: very large models pre-trained on massive datasets and then adapted to specific downstream tasks with relatively small amounts of task-specific data.[36,37,43] In medical imaging, medical natural language processing, and surgical AI, Transformer-based architectures have increasingly outperformed

LSTM and convolutional predecessors, particularly when global context, multi-modal integration, or long sequences are critical for temporal forecasting.[37,39,40,43] For example, in medical image analysis, vision Transformers utilize self-attention mechanisms to model global spatial relationships across the whole image, allowing them to capture complex and long-range anatomical context missed by other deep learning networks.[40] The utility of foundation model strategies, i.e., pre-training on large, heterogeneous data and then adapting to specialized tasks, is one of the principal reasons Transformer-based architectures are purposive in surgical AI-based training where realistic patient data are limited.[37,39,40,43]

LSTMs can outperform Transformers when datasets are small and sequences are short; Transformers outperform LSTMs when datasets are larger and predicting future behavioral patterns based on long-range multi-instrument interactions are critical.[36-38] A sequence refers to the time-series of kinematic metrics recorded during a subpial resection trial. A short sequence may represent a relatively shorter time-series, such as only a few seconds, of local instrument motion. In contrast, a long sequence may span a longer procedural timeframe, such as pial incision, bipolar elevation, aspirator resection, and instrument transitions. Hybrid architectures that combine the two: using the Transformer's attention to capture global context via large-scale statistical pattern modelling and the LSTM's recurrent processing to capture fine-grained local temporal dynamics, have been explored in adjacent domains including student performance analysis and speech emotion recognition.[41,42] However, being relatively new, Transformers remain virtually absent in the realm of surgical education and surgical AI/intelligent tutoring systems. Thus, whether any of these AI architectures is optimal for an intelligent tutoring system trained on data from a realistic ex vivo surgical model to forecast future behaviors across training years is an open empirical question.[34,36-38,41,42]

The ex vivo setting sharpens the architectural question for two reasons. First, as already noted, the absence of force and safety-related metrics in a non-perfused ex vivo model shifts the feature set toward kinematic (motion) metrics (velocity, acceleration, jerk, inter-instrument distance).[28,34] These kinematic metrics vary on multiple timescales: short, intermittent bursts of acceleration during pial incision, sustained coordination during aspiration, and procedure-wide complex dynamics.[28,34] Second, an ex vivo procedure using the actual operative instruments is more complex than a VR subpial resection as it implements three instruments rather than two, spans a longer duration, embodies three-dimensional haptic and ergonomic adaptations, and includes a preparation phase and a suction/aspirator resection phase that are kinematically distinct.[28,34] A predictive model that captures global patterns across instruments and phases may therefore have an advantage over a model that processes the data sequentially.[28,34,36]

These considerations motivate the architectural comparison between the VR-validated LSTM, a Transformer, and two of their hybrid combinations: an LSTM–Transformer (Hybrid 1), and a Transformer–LSTM (Hybrid 2), trained under identical preprocessing and evaluation methods on the same ex vivo calf brain subpial corticectomy data.[28,36,38,41,42] The goal is to identify the best-performing AI architecture for an EV-ICEMS but also to establish a template for architectural choices in future intelligent tutoring systems trained on realistic surgical data, applicable broadly across surgical AI systems.[28,36,38,41,42]

Toward the Intelligent Operating Room: EV-ICEMS as the Bridge

By learning individualized performance patterns, transformer-based models may help future intelligent tutoring systems personalize feedback by identifying trainee-specific behavioral patterns over time.[27-30,36,43] The integration of LLMs further extends this paradigm, enabling natural language feedback, interactive coaching, and explainable guidance embedded directly

within the training loop.[43,44,114] The next evolution of surgical AI is being driven by Big Data ecosystems that fuse high-dimensional inputs—operative video, instrument kinematics (velocity, acceleration, jerk) and multi-instrument coordination patterns (behavioral patterns), and even emerging proxies of cognitive and emotional states—into unified learning frameworks.[44,65-71,90-93,115,116] Within this paradigm, foundation models trained on large-scale multimodal surgical datasets, including VR platforms and ex vivo simulations where data can be more safely and abundantly generated, enable continuous monitoring, fine-grained skill assessment, and adaptive learning at scale.[43,44,90-93,114]

The long-term goal of the research program to which this thesis contributes is an intelligent OR, i.e., an intraoperative AI system in a live OR environment which continuously monitors and assesses surgical performance, flags deviations as early risk signals from expert benchmarks, and prevents errors in real-time.[44,90–93,115] The problem between where the field currently stands, validated AI tutoring in VR, and that goal is substantial. Unlike VR systems, which rely on simulator-specific physics such as simulated bleeding and tissue injury, the ex vivo model uses real operative instruments and generates instrument-tracking data closer to what could be captured during live surgery.[33,34,64] An intelligent tutoring system built and validated on ex vivo data is therefore a methodological bridge toward an intraoperative intelligent system.[33,34,44,90–93] Its construct and predictive validity across training levels are the first evidence that the approach can differentiate surgical expertise in a realistic ex vivo environment and the foundation for subsequent studies that will extend the approach toward live operative data.[45-47,90-93,115]

Ethical translation to an intelligent OR will, of course, require work well beyond the scope of this thesis. In this context, *primum non nocere* — safety and harm prevention — must remain the hallmarks of human–AI synergy in the intelligent operating room. Privacy, patient agency,

accountability, and robust AI safety frameworks are central concerns in any deployment of continuous AI assessment during patient care.[48,91,92,110,115,116] The most useful contribution this thesis can make to that longer-term program is methodological: to determine which deep learning architecture motivated by prior ICEMS work, best supports an intelligent tutoring system in a realistic ex vivo surgical environment, and to provide rigorous construct and predictive validity evidence for the resulting system.[34,36,38,45–47]

Four architectures were explored herein: LSTM, Transformer, and two hybrids, namely LSTM–Transformer, and Transformer–LSTM. Alternative time-series modelling AI architectures, including other recurrent neural networks, convolutional networks, and foundation-model approaches, may also be relevant to future surgical AI systems.[28,40,43] Emerging foundation models rely on Transformer architectures as their core computational backbone.[43] However, prior ICEMS work established an LSTM architecture as the validated model for continuous expertise assessment in VR surgical simulation. Our preliminary model exploration also supported focusing the present EV-ICEMS comparison on whether this LSTM architecture should be replaced or augmented by Transformer-based architectures in the ex vivo setting, given Transformers' parallel processing across time-series data.[28,36,38,43] Further, hybrid architectures were evaluated because LSTM–Transformer hybrid approaches have been shown to be effective in educational and affective computing domains, including student performance analysis and speech-emotion recognition.[41,42]

THE STUDY OBJECTIVES

The primary objective of this thesis was to determine which deep learning architecture, trained on kinematic data from a calf brain subpial corticectomy model, best optimizes the development of the Ex Vivo Intelligent Continuous Expertise Monitoring System (EV-ICEMS).

The secondary objective was to assess the construct and predictive validity of the model architectures across training levels, including novices, juniors, seniors, and experts.

THE STUDY HYPOTHESIS

We hypothesized that a Transformer or Transformer-based hybrid architecture would outperform an LSTM-based architecture for EV-ICEMS development due to its attention layer's ability to capture global, long-range relationships among instruments during complex surgical procedures. We further hypothesized that Transformer-containing architectures would demonstrate stronger construct and predictive validity than the LSTM alone.

THE STUDY

TITLE: Transformer Architecture Optimizes Intelligent Tutoring Systems for Surgical Training

AUTHORS: Abicumaran Uthamacumaran, BSc[1,2]; Abdulrahman Almansouri, MBBCh, MSc[1,3]; Bianca Giglio, MSc[1,4]; Matheus Ballestero, MD, PhD[1,5]; Nour Abou Hamdan, MSc[1]; Recai Yilmaz, MD, PhD[1,6]; Louis Collins, PhD[4]; Houssem-Eddine Gueziri, PhD[4,7]; José A. Correa, PhD[8]; Rolando F. Del Maestro, MD, PhD[1,3]

AFFILIATIONS:

1. Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
2. Department of Surgical and Interventional Sciences, McGill University, Montreal, Quebec, Canada
3. Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada
4. Department of Biomedical Engineering, McGill University, Montreal, Quebec, Canada
5. Department of Medicine, Federal University of São Carlos, São Carlos, Brazil
6. Children's National Medical Center, Division of Neurosurgery and Pediatrics, Washington, DC, United States of America
7. Université TÉLUQ, Montreal, Quebec, Canada
8. Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada

CORRESPONDING AUTHOR: Abicumaran Uthamacumaran, BSc, Neurosurgical Simulation and Artificial Intelligence Learning Centre, Montreal Neurological Institute and Hospital, 300 Rue Léo-Pariseau, Ste 2210, Montreal, QC, Canada H2X 4B3 (abicumaran.uthamacumaran@mail.mcgill.ca)

This manuscript was submitted for review to Scientific Reports on May 8, 2026.

Abstract

Translation of intelligent tutoring systems to operative settings remains challenging due to the sequential nature of procedures and data limitation. The deep learning architecture which best optimizes improvement of these educational systems for surgical training remains unknown. This investigation therefore evaluated Long Short-Term Memory (LSTM) networks, Transformers, and hybrid architectures to develop individual calf brain Ex Vivo Intelligent Continuous Expertise Monitoring Systems (EV-ICEMS). This case series study included 47 participants, novices, junior and senior residents/fellows along with neurosurgeons who performed three subpial ex vivo corticectomies on the calf brain model. Primary outcomes included composite expertise scores (–1.00 [novice] to 1.00 [expert]) and assessment of each model's construct and predictive validity. The Transformer and LSTM-Transformer demonstrated better construct validity than other architectures but only the Transformer differentiated seniors from juniors (mean difference, 0.59; 95% CI, 0.23 to 0.96; P = .004). The Transformer demonstrated superior predictive validity (R^2 = 0.424; MSE = 0.168; MAE = 0.365) with the steepest score-year slope (0.150; 95% CI, 0.074 to 0.225). This study outlines a deep learning methodology to differentiate

surgical expertise levels and optimize the development of intelligent tutoring systems for surgical training in realistic operative environments.

Keywords: surgical simulation; deep learning; intelligent tutoring system; Transformer; long short-term memory; surgical expertise

INTRODUCTION

Artificial intelligence (AI) is changing surgical education by quantifying technical skill execution and generating novel performance metrics, enabling competency-based assessment and personalized feedback.[98,99] Within this framework, data-driven intelligent tutoring systems are being utilized in surgical education.[25-29,105,108] These platforms focus on developing AI-enhanced curricula, providing quantitative trainee assessment for effective real-time feedback essential for learner engagement, training, and error mitigation.[26,27,29,31,32,108]

The NeuroVR (CAE Healthcare) simulation platform captures detailed instrument movements and interactions, enabling bimanual psychomotor skill tracking and automated coaching.[26,29,31,32,64] This platform was used to develop the Intelligent Continuous Expertise Monitoring System (ICEMS), an intelligent tutoring system based on a Long Short-Term Memory (LSTM) deep neural network.[25,28] The ICEMS assesses bimanual surgical skills at 0.2-second intervals based on AI-derived metrics and provides trainees with personalized, actionable verbal feedback.[28] The ICEMS has been validated and outperformed human instructors in a series of randomized controlled trials (RCTs) and crossover studies.[26,27,29,30] AI-augmented

personalized human educator instruction resulted in improved surgical performance and skill transfer, and fewer errors compared with ICEMS instruction alone, highlighting the need for the development of more proficient intelligent tutoring systems.[27,105]

Deep learning models like LSTM networks only process data sequentially but excel at capturing temporal dynamics of performance, while transformers process data in parallel which aids in understanding context and complex data relationships.[39,40] LSTM networks can outperform Transformers when data volumes are small and involve short sequence lengths, while Transformers outperform LSTM applications when long-range dependencies are critical and datasets are large.[37,97] Transformer-LSTM hybrid models leverage the transformer's attention capabilities and the LSTM's temporal modeling.[41] The use of Transformer-LSTM hybrid architectures is being explored to assess student academic performance and speech emotional recognition when large datasets are available.[41,42] It is unknown whether LSTM networks or Transformers alone or their combination can improve the function of intelligent tutoring systems utilized for surgical training in which limited datasets are common.

These AI technologies have also not been developed for or tested with more biologically realistic models in which surgical instrument tracking has been utilized.[33-35,75] Ex vivo models offer anatomical realism in controlled, patient risk-free operative environments, allowing for the development of novel metrics applicable to the use of specific operative instruments across many surgical simulation platforms.[33,34,49] The subpial corticectomy is a critical bimanual task that neurosurgical trainees must master for the resection of brain tumors and epileptic foci.[84,85] An ex vivo calf brain subpial corticectomy model developed by our group has demonstrated face and

content validity, as well as instrument tracking capabilities.[34] One goal of ex vivo investigations is the development of data-driven intelligent tutoring systems for surgical education.

This study aimed to determine which deep learning architecture, trained on data from a calf brain subpial corticectomy model, would demonstrate superior construct and predictive validity, thus optimizing the development of the Ex Vivo Intelligent Continuous Expertise Monitoring System (EV-ICEMS). We hypothesized that a Transformer-based architecture would outperform an LSTM model due to its attention layer's ability to capture global, long-range relationships such as instrument usage in complex surgical procedures.[37,97]

RESULTS

Participants and Dataset

A total of 47 participants were enrolled and allocated a priori to 4 groups. Groups included 14 medical students (novices), 14 post-graduate year (PGY)-1–5 neurosurgical residents (juniors), 4 PGY-6 residents and 7 fellows (n = 11) (seniors), and 8 neurosurgeons (experts). Participant demographic data are outlined in **Table 2**. Each of the 47 participants performed 3 subpial corticectomies (**Figure 2**). However, due to tracking information loss, only 136 trials were available for review (**eTable 1** in Supplement 1). Participant grouping was based on training level and clinical expertise. Although Table 2 reports specialty background for fellows and neurosurgeons, the pre-task questionnaire[34] captured broad surgical interest items for medical students and self-reported prior technical experience related to subpial resection in epilepsy and

tumor cases with estimated counts for some participants. However, these variables were heterogeneous, incompletely standardized, and not incorporated as predefined covariates in the present study. They could influence motivation, task familiarity, and composite score variability, and should therefore be prospectively standardized in future studies.

Model Performance

The Transformer and Hybrid 1 models demonstrated the best construct validity compared with other AI model architectures (**Figure 3**A, C, E, and G). However, the Transformer model was the only model to differentiate seniors from juniors (**Figure 3**C), while only Hybrid 1 differentiated experts from seniors (**Figure 3**E). The Transformer demonstrated superior predictive validity across training years (**Figure 3**D).

Expertise Group Discrimination

The Transformer and Hybrid 1 demonstrated strong construct validity. For the Transformer, mean scores were 0.69 (95% CI, 0.47 to 0.91) for experts, 0.41 (95% CI, 0.25 to 0.56) for seniors, –0.18 (95% CI, –0.53 to 0.16) for juniors, and –0.80 (95% CI, –0.99 to –0.61) for novices. Significant pairwise differences were observed between experts and juniors (mean difference, 0.87; 95% CI, 0.49 to 1.26; $P < .0001$), experts and novices (mean difference, 1.49; 95% CI, 1.22 to 1.76; $P < .0001$), seniors and juniors (mean difference, 0.59; 95% CI, 0.23 to 0.96; $P = .004$), seniors and novices (mean difference, 1.20; 95% CI, 0.97 to 1.43; $P < .0001$), and juniors and novices (mean difference, 0.61; 95% CI, 0.23 to 0.99; $P = .001$).

For Hybrid 1, mean scores were 0.75 (95% CI, 0.52 to 0.97) for experts, 0.25 (95% CI, –0.05 to 0.56) for seniors, –0.09 (95% CI, –0.39 to 0.21) for juniors, and −0.80 (95% CI, –0.92 to –0.68) for novices. Significant differences were observed between experts and seniors (mean difference, 0.49; 95% CI, 0.14 to 0.85; $P = .045$), experts and juniors (mean difference, 0.84; 95% CI, 0.49 to 1.19; $P < .0001$), experts and novices (mean difference, 1.55; 95% CI, 1.31 to 1.79; $P < .0001$), seniors and novices (mean difference, 1.05; 95% CI, 0.74 to 1.37; $P < .0001$), and juniors and novices (mean difference, 0.71; 95% CI, 0.40 to 1.02; $P < .001$). Differences between seniors and juniors were not statistically significant.

Hybrid 2 demonstrated more modest group separation. Mean scores were 0.56 (95% CI, 0.21 to 0.90) for experts, 0.08 (95% CI, –0.23 to 0.39) for seniors, –0.05 (95% CI, –0.31 to 0.21) for juniors, and –0.60 (95% CI, –0.78 to –0.41) for novices. Significant differences were observed between experts and juniors (mean difference, 0.60; 95% CI, 0.20 to 1.01; $P = .01$), experts and novices (mean difference, 1.15; 95% CI, 0.78 to 1.53; $P < .0001$), seniors and novices (mean difference, 0.68; 95% CI, 0.33 to 1.03; $P = .001$), and juniors and novices (mean difference, 0.55; 95% CI, 0.24 to 0.86; $P = .006$).

LSTM showed the weakest construct validity. Mean scores were 0.42 (95% CI, 0.11 to 0.74) for experts, 0.15 (95% CI, –0.13 to 0.42) for seniors, –0.11 (95% CI, –0.29 to 0.08) for juniors, and –0.52 (95% CI, –0.83 to –0.21) for novices. Significant differences were observed between experts and juniors (mean difference, 0.53; 95% CI, 0.19 to 0.87; $P = .03$), experts and novices (mean difference, 0.94; 95% CI, 0.54 to 1.35; $P < .0001$), and seniors and novices (mean

difference, 0.67; 95% CI, 0.28 to 1.06; *P* = .002).

Quantifying Skills Across Training Years

For predictive validity across training years, the Transformer achieved an $R^2 = 0.424$ with MSE = 0.168 and MAE = 0.365 on previously unseen surgical training data, followed by LSTM (R^2 = 0.234; MSE = 0.105) and Hybrid 1 ($R^2 = 0.231$; MSE = 0.193), while Hybrid 2 showed the weakest predictive performance ($R^2 = 0.105$; MSE = 0.177) (**Figures 3**B, D, F, and H). The estimated slope for score versus year of training was highest for the Transformer (0.150; 95% CI, 0.074 to 0.225), compared with LSTM (0.076; 95% CI, 0.017 to 0.136), Hybrid 1 (0.103; 95% CI, 0.022 to 0.183), and Hybrid 2 (0.061; 95% CI, –0.016 to 0.139).

Comparison of tool-metric pairs demonstrated that participant utilization involving microscissors and bipolar kinematic-based metrics distinguished Transformer and Hybrid 1 models from other architectures (**Figure 4**). For the Transformer, all three microscissors kinematic-based metrics (velocity [mean difference, 0.59; 95% CI, 0.10 to 1.09; *P* = .01], acceleration [mean difference, 0.49; 95% CI, 0.0003 to 0.97; *P* = .0498], and jerk [mean difference, 0.52; 95% CI, 0.003 to 1.00; *P* = .035]) were significantly different between seniors and juniors. For Hybrid 1, two microscissors kinematic-based metrics (velocity [mean difference, 0.58; 95% CI, 0.15 to 1.02; *P* = .005] and jerk [mean difference, 0.57; 95% CI, 0.076 to 1.00; *P* = .02]) were significant. For Hybrid 1, two bipolar kinematic-based metrics (velocity [mean difference, 0.55; 95% CI, 0.08 to 1.01; *P* = .02] and acceleration [mean difference, 0.48; 95% CI, 0.01 to 0.95; *P* = .04]) were significantly different between seniors and juniors. Only bipolar velocity [mean difference, 0.56;

95% CI, 0.10 to 1.02; $P = .01$]) was significantly different utilizing Transformer architecture (**Figure 4**; **eTable 3** in Supplement 1).

DISCUSSION

To the authors' knowledge, this is the first study to assess whether alternative deep learning architectures optimize surgical training utilizing actual operative instruments in realistic surgical settings. Consistent with our hypothesis, this investigation demonstrates that the Transformer architecture trained on data from a calf brain subpial corticectomy model provides strong evidence for construct validity and superior predictive validity, thus optimizing EV-ICEMS development. A recent review concerning simulation training in healthcare fields revealed that only a small fraction of training systems had studies supporting construct and predictive validity, highlighting the need to develop more relevant simulation training tools.[45,50] This investigation outlines a deep learning methodology which may be useful in optimizing the development of intelligent tutoring systems for the assessment and training of procedural interventions. Future studies should incorporate multimodal sensing including force, haptics, and safety metrics, to enable more comprehensive assessment of surgical performance. Integration of cognitive and emotional dimensions, such as cognitive load, may further enhance holistic surgical education.[48,51,116]

In VR-based surgical education, an LSTM-based architecture has been effectively used to build an intelligent tutoring system that assesses operative performance and mitigates errors.[26-29] The

present study's Transformer model demonstrated higher predictive power (R2 = 0.424) than the LSTM-based architecture in our previous VR study (R2 = 0.277).[28] Both Transformer-based models (Transformer and Hybrid 1)—where the Transformer performed the final expertise regression—demonstrated the strongest construct validity, capturing significant expertise-related differences driven predominantly by microscissors and bipolar kinematic-based metrics. In contrast, LSTM primarily captured aspirator-based differences, which were less discriminative. This may reflect the Transformer's attention-based architecture, which captures long-range temporal dependencies and global coordination patterns across multi-instrument metrics.[36] LSTM-based architectures focus on sequential processing, which is common in operative procedures.[38] However, they may be unable to provide context related to the whole series of instrument movements involved in complex operative procedures.[38]

In our study, the deep learning architectures studied demonstrated differentiation across expertise groups, with experts achieving the highest composite scores across models. This suggests that technical mastery of subpial corticectomy continues to develop beyond residency and fellowship, consistent with prior VR study findings.[27-29] However, at the composite score level, only the Transformer significantly differentiated seniors from juniors, outlining that its principal advantage was predictive validity (Figure 3). This is further supported by the steeper slope observed in the Transformer (0.150) compared with Hybrid 1 (0.103), indicating greater incremental improvement in skill assessment across training years. Additionally, the Transformer demonstrated the highest area under the curve (AUC) and test accuracy on previously unseen novice and expert data during the train-validation-test split (eTable 5 in Supplement 1).

Our Transformer-based EV-ICEMS provides surgical educators and learners with transparency and explainability essential to optimizing learning outcomes.[25] In this study, microscissors and bipolar kinematic-based metrics were the most important determinants of surgical expertise. This would suggest that analysis of complex, individual instrument metrics is essential to developing intelligent tutoring systems for surgical procedures. Educators and learners provided with this information can more effectively develop AI-based curricula to maximize learner engagement and performance.[54] Consistent with our previous studies involving a subpial resection VR model, bipolar-to-aspirator instrument separation distance was not a primary determinant of expertise.[27,29] In these VR investigations, tissue injury secondary to instrument force application along with bleeding risk, were critical in RCTs and crossover studies to improve trainee acquisition of surgical skills and error reduction.[27,29,30] We were unable to assess these metrics in the ex vivo model used in this study.[34] Developing methodologies to quantitatively assess force application, tissue injury, and bleeding risk in the ex vivo model will substantially improve our capacity to develop more capable intelligent tutoring systems for training.[117]

Ex vivo surgical simulation platforms provide biologically realistic environments for training bimanual psychomotor skills while enabling evaluation of model validity and preserving standardized, reproducible task assessment.[34] These models provide explainable, quantitative metrics and high tissue fidelity biofeedback.[34] By generating large datasets of "skilled" and "less skilled" surgical instrument performance, they improve the granularity of participant classification and the identification of novel metrics that better characterize expertise.[31]

The EV-ICEMS can be used to develop platforms capable of providing real-time, continuous,

action-oriented verbal feedback for training and error mitigation.[28] The implementation of these systems may allow surgical educators to enhance simulation-based and AI-enhanced surgical curricula.[26-29,54] The development and validation of equivalent AI-powered intelligent tutoring systems in the human operating room (OR) are the long-term goals of this study.

LIMITATIONS

The system does not capture the spectrum of competencies required in dynamic OR environments, including real-time interactions with human educators and other healthcare professionals. The absence of safety-related metrics (e.g., tissue injury and bleeding), instrument force application along with emotional and cognitive components of learning limits the ability to assess operative risk and patient safety.[48,51,116] The cohort size was small and drawn from three Quebec institutions, limiting applicability to other learners and experts. This study reports results for subpial resections in an ex vivo calf brain model and generalization to other surgical procedures will need further investigation. Prospective studies will be needed to support comparative inference regarding trainee performance across transformer-based EV-ICEMS models using multi-institutional randomized trials with larger sample sizes.
Ethical translation to an intelligent OR will require robust AI safety and responsible AI frameworks prioritizing privacy, patient agency, and accountability.[33]

The integration of Transformer-based multimodal embeddings with computer vision and large language model–based feedback systems may further enhance the utility of intelligent tutoring systems.[27,44,93,105,114] Foundation models are Transformer-based AI systems initially trained on

massive datasets and subsequently adapted to specific tasks using more limited data.[43] The utility of these models in surgical training should be further explored.

AI-enhanced curricula need to incorporate adaptive training elements that can continuously adapt to each learner's skill set proficiencies and learning trajectory to maximize capacity for task mastery.[118] Importantly, AI-augmented instruction should remain embedded within human-led mentorship, reinforcing that these systems are primarily designed to augment student learning, not replace, the surgical educator.[27,105]

CONCLUSION

In conclusion, this study outlines a deep learning methodology to differentiate surgical expertise levels, enabling the optimization of intelligent tutoring systems. The Transformer architecture trained on data from a calf brain subpial corticectomy model demonstrated superior predictive validity for simulation training of bimanual procedural interventions in realistic surgical settings.

METHODS

Participants

Medical students, neurosurgical trainees, and neurosurgeons from Quebec universities were enrolled in this case series from September 2022 to May 2023. Data were recorded at a single time point without follow-up. Each participant performed three different simulated subpial corticectomies as previously described.[34] Standardized verbal and written task instructions were provided before the simulation, and participants were able to ask clarification questions before and during the task. Verbal clarification of task instructions could be provided in English or French when required. Since participants were recruited in Quebec, consent forms were available in English and French, and the consent discussion was conducted in the participant's preferred language when required. One participant selected the French consent form. This study was conducted in compliance with the principles of the Declaration of Helsinki and was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry Ethics Board (Approval No. IRB00010120). Participants signed an approved consent form before trial commencement and were not made aware of the trial's purpose nor assessment metrics at any point. This study follows the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT–AI) guidelines[115] and the Machine Learning to Assess Surgical Expertise checklist.[52]

Ex Vivo Brain Model and Continuous Instrument Monitoring

Because of their morphological similarity to the human pediatric brain, as well as their availability, low cost,[33-35,75] and usefulness for microsurgical training,[34] calf brains were used in this study. Fresh ex vivo calf brains with similar weight, structure, and well-defined gyri were obtained postmortem from a local butcher through the commercial food supply chain. No live animals were used or euthanized for research purposes. Development and assessment of the ex vivo calf brain simulation scenarios followed best practice guidelines for ex vivo animal brain models in neurosurgical education using the "ex vivo brain model to assess surgical expertise" (EVBMASE) checklist.[33]

Fresh calf brains were positioned within a human skull model utilizing an off-midline craniotomy and operated on under an OPMI Pico surgical microscope.[34] Microscissors, bipolar forceps, and an ultrasonic aspirator (Stryker) were equipped with customized PLA-mounted fiducial markers (Figure 1).[34] These markers enabled real-time 3D tracking using a dual-camera infrared optical system (FusionTrack 500, Atracsys; Polaris, Northern Digital Inc.). Tracking data were streamed into 3D Slicer (version 5.0.3) through the PLUS/OpenIGTLink acquisition pipeline.

Data Acquisition and Preprocessing

The microscissors, bipolar, and aspirator tip positions were measured on average every 0.1 seconds (10 Hz) during task duration, allowing derivation of three kinematic metrics (velocity, acceleration, and jerk) for each instrument and the bipolar–aspirator separation distance. This

0.1-second sampling interval refers to the present ex vivo instrument-tracking system.[34] By contrast, the 0.2-second interval described for prior ICEMS work refers to the earlier VR-based ICEMS platform.[28] The kinematic time-series were denoised and smoothed using a fourth-order low-pass Butterworth filter. These performance metrics were selected because they involve coachable bimanual surgical psychomotor behaviors. Mean instrument detection rates were 79% (microscissors), 88% (bipolar), and 74% (aspirator) with no significant differences between groups. Full computational formulas for all derived metrics are provided in **eTable 2** in Supplement 1.

Outcome Label and EV-ICEMS Score

As previously described by Yilmaz et al., the AI algorithms were trained on data from expert (neurosurgeon) and novice (medical student) trials only.[28] The senior and junior trainee performance was unseen to the models and was used for post-AI training validation.[28] AI models assessed performance at 0.1-second intervals and calculated expertise scores ranging from –1 (novice) to 1 (expert). Participant-level mean scores were computed by averaging trial-level predictions.

Deep Learning Architectures and Training

Four deep learning architectures were evaluated for sequential skill modeling: (1) LSTM, (2) Transformer, (3) LSTM-Transformer (Hybrid 1), and (4) Transformer-LSTM (Hybrid 2). All models were trained under identical preprocessing and evaluation frameworks using 70/15/15

train/validation/test splits (**Figure 2**).

All models processed instrument time-series sequences per participant and per trial, including kinematic metrics (velocity, acceleration, and jerk) and bipolar-to-aspirator distance, as their input layer. Details concerning model training are available in **eFigure** in Supplement 1.

Statistical Analysis and Model Evaluation

ICEMS model performance metrics were predictive accuracy (R^2, coefficient of determination), mean squared error (MSE), and mean absolute error (MAE) on previously unseen test data. To assess construct and predictive validity across training strata, the average performance metrics (ICEMS scores) were compared across 4 expertise levels (experts, seniors, juniors, and novices) using one-way analysis of variance (ANOVA). Assumptions of ANOVA model errors, including normality and homogeneity of variance, as well as the presence of outliers or influential observations were assessed by graphical examination of model residuals. When warranted, post hoc pairwise comparisons of mean differences were performed, and *P* values were adjusted for multiple testing using the Tukey method.

All statistical tests of hypotheses were two-sided and performed at the 5% significance level. Statistical analyses were conducted in R (R Foundation).[119] In complementary analyses, we regressed predicted score against year of training using ordinary least squares with 95% confidence bands, reporting the R2 as the main fit statistic, with MSE and MAE as error metrics. Hyperparameter tuning was guided by parameters validated in the previously established LSTM-

based ICEMS framework within a VR setting[28] and matched among the 4 models.

DATA AVAILABILITY STATEMENT: The dataset is available from the authors on a reasonable request.

CODE AVAILABILITY STATEMENT: The codes used in this study are available from the authors on a reasonable request.

ARTIFICIAL INTELLIGENCE USE STATEMENT

No large language models or generative artificial intelligence tools were used in the preparation of this manuscript.

THESIS SUMMARY

Discussion

This thesis presents the development and validation of the EV-ICEMS, a deep learning framework trained on kinematic and behavioral performance data captured during a calf brain subpial corticectomy simulation. The central aim was to determine which deep learning architecture most accurately differentiates surgical expertise levels from realistic ex vivo instrument tracking data, and to examine what the learned representations reveal about surgical expertise and its underlying teachable metrics. This study is the first to benchmark Transformer vs. LSTM architectures for surgical skill assessment using real operative instruments on a realistic ex vivo brain model.

The ex vivo model employed optical tracking via fiducial markers to capture high-resolution, real-time three-dimensional kinematic data from three surgical instruments: microscissors, bipolar forceps, and aspirator. Ten performance metrics were extracted per instrument across four surgical expertise groups: novices (medical students), juniors (PGY 1–5), seniors (PGY 6 and fellows), and experts (neurosurgeons). These instrument motion-based metrics, including velocity, acceleration, jerk, and inter-instrument tip separation distance, were fed into four model architectures that each produced a composite expertise score on a continuous scale from −1 (novice) to 1 (expert). Consistent with our hypothesis, the Transformer generated strong construct validity and superior predictive validity.

The findings of this thesis advance the field of surgical AI and intelligent tutoring systems in three important ways. First, they establish that Transformer-based architectures outperform

sequential LSTM models in the realistic, limited-sample ex vivo model (Figure 3). Second, they reveal that microscissors and bipolar kinematic metrics are the most critical predictors of expertise in this bimanual corticectomy task, suggesting that the Transformer architecture detects expertise-specific technique patterns embedded within instrument motion (Figure 4; Appendix C). These findings should be interpreted as task- and model-specific. These metrics were the most discriminative signals in this standardized, single-operator ex vivo subpial corticectomy task, and not universal determinants of neurosurgical expertise. Third, they demonstrate that the composite metric score captures emergent behavioral dynamics that transcend any single kinematic measure, alluding toward data-driven predictors of expert surgical performance with profound implications for the design of next-generation intelligent tutoring systems and eventually an intelligent OR.

This work must be understood alongside the broader program of research conducted at the Neurosurgical Simulation and Artificial Intelligence Learning Centre at McGill University. These findings underscore that while AI-driven performance data enhances learning when integrated with human instruction, the emotional and cognitive dimensions of the learner experience must be carefully considered. LLM-based feedback systems that can contextualize and personalize explanations may be key to managing these affective-cognitive responses and enabling the full translational potential of EV-ICEMS in surgical curricula. [27,30,43,44,65-71,105,114,116]

Microscissors and Bipolar as Key Predictors of Expertise

Tool-metric pairwise comparison analyses revealed that the Transformer and Hybrid 1 generated the strongest patterns of statistically significant between-group differences (Figure 4). For the Transformer, all three microscissors kinematic metrics (velocity, acceleration, and jerk) successfully separated senior from junior residents, while Hybrid 1 (LSTM-Transformer)

identified microscissors velocity and jerk, and bipolar velocity and acceleration as the key discriminators between these groups. These findings reflect the technique structure of the subpial corticectomy. As shown in Appendix C, the attention layers which weight input features by their relevance to the model's prediction, and integrated gradients of the trained Transformer identified jerk as a salient determinant of expertise across instruments. Jerk, as the rate of change of acceleration, quantifies movement smoothness, with elevated jerk reflecting jittery, less controlled motions. Expert surgeons generate smooth, deliberate movements with minimized abrupt shifts, a pattern coachable through slow and deliberate practice, reduction of sudden direction changes, and smooth velocity and acceleration transitions.

A critical design decision was the inclusion of all ten metrics simultaneously. Analyses confirmed that this full metric set yielded the highest regression R^2 and lowest prediction error, removing any individual metric, even bipolar-to-aspirator tip separation distance, degraded group separation and predictive validity. This illustrates a Gestalt principle inherent to complex systems: the ‘whole is greater than the sum of its parts’. Every instrument–metric pair plays a synergetic role, encoding unique behavioral information. The latent representations learned by the attention layers encode emergent features of behavioral dynamics, including the nonlinear coordination patterns of multi-instrument usage across precise temporal windows. This also explains why bipolar-to-aspirator tip separation distance was not statistically meaningful on its own yet appeared as a salient feature in the Transformer's attention layers and integrated gradients (Appendix C). For learners and educators, EV-ICEMS should emphasize explainable, instrument-specific coaching, particularly focused on microscissors and bipolar kinematics, to guide technique development from competence toward mastery in in AI-enhanced surgical curricula.

Methodological Innovations

The specific architectures evaluated here represent a snapshot of a rapidly evolving AI landscape. While Transformers and recurrent neural networks remain the dominant paradigms in current data-driven predictive models, future innovations in AI architectures will bring new representational capacities to surgical expertise prediction. What endures is our approach itself: kinematic time-series data from multi-instrument ex vivo tracking contains nonlinear, emergent patterns (synergy) of behavioral dynamics that sufficiently powerful deep learning architectures capable of learning parallel, context-sensitive global representations, like attention mechanisms of Transformers, can decode into accurate expertise predictions and intelligent tutoring systems. The latent space of a well-trained Transformer encodes not merely motion, but behavioral coordination, and the distributed temporal signature of surgical expertise across instruments as dynamical systems.

Equally important is the signal engineering that precedes modeling. As detailed in Appendix B, a fourth-order Butterworth low-pass filter was identified as the optimal denoising solution: in data preprocessing, attenuating high-frequency noise while preserving the behaviorally relevant frequency content of instrument motion. The quality of the learned representation is bounded by the quality of the input signal. No architecture can recover emergent behaviors destroyed by noise fluctuations. The fourth-order Butterworth filter combined with the Transformer architecture are therefore complementary innovations: one preparing the signal for learning, the other extracting the expertise signatures embedded within the time-series metrics.

Limitations and Future Directions

Several important limitations of the current work must be acknowledged. First, and most critically, this ex vivo model does not yet quantify safety-related metrics such as tissue injury risk, bleeding risk, or instrument force application.[28-30,34] These dimensions are central to operative safety and constitute key targets for next-generation EV-ICEMS development. One promising avenue is the incorporation of simulated tumor models, such as gel-based tumor grafts previously developed by our group, into the ex vivo preparation.[35,75] Separately, a bleeding vessel simulation could be integrated into the ex vivo model to allow measurement of hemorrhage control performance by utilizing perfusion of intact carotid arteries in this model. Most importantly, future iterations should incorporate force sensing and haptic feedback instrumentation to measure instrument-tissue interaction forces, providing richer multimodal data for EV-ICEMS training. Specifically, the integration of force sensors, surgical robotics, and flexible force-sensing arrays onto instrument handles could potentially circumvent this barrier, capturing finer dimensions of operative performance that kinematic data alone cannot encode.

Second, the EV-ICEMS as implemented does not capture the full complex and adaptive spectrum of competencies required in the dynamic, adaptive operating room environment. Surgical expertise encompasses not only bimanual psychomotor skill but also team-based learning and communication, situational awareness, adaptive decision-making, and dynamic interaction with human educators and a multidisciplinary team. The present task was performed as a standardized single-operator simulation[34], whereas many real-world neurosurgical procedures are performed under the microscope by a lead surgeon and an assistant or second surgeon. In some cases, bimanual coordination is distributed across the team: one surgeon may hold the bipolar while another manages complementary techniques such as suction, retraction, microscissors, hemostasis,

or tissue mobilization. Therefore, the present ex vivo model does not capture surgeon-assistant coordination, instrument handoffs, verbal and nonverbal communication, or collaborative crisis management.[48,116] The absence of these team-based dimensions limits the ecological complexity of EV-ICEMS as a comprehensive competency framework, and future systems must aspire toward multi-dimensional assessment. This team-based dimension is central to surgical competence and to the instructor-trainee relationship; its absence from current AI-based surgical curricula represents an important limitation.[48,116,120]

Third, this model represents only one technically critical segment of a neurosurgical procedure rather than the full operative pathway.[34] Before a subpial resection begins, surgeons must select the operative approach, position the patient and head, plan pin placement, open the skin and cranium, decide whether cerebrospinal fluid diversion or retractors are needed, choose the surgical corridor, navigate eloquent cortex and vascular anatomy, interpret intraoperative pathology, decide the safety margins of resection, and close the wound to minimize complications such as cerebrospinal fluid leak, wound infection, and postoperative hemorrhage.[2,3,27,28,34,55,84,85] These cognitive, anatomical, and strategic decisions may challenge residents and fellows as much as instrument motion at the resection cavity. The current EV-ICEMS should therefore be interpreted as a model of bimanual psychomotor performance during a defined subpial corticectomy phase, not as a complete assessment of neurosurgical competence. Future EV-ICEMS development should incorporate staged scenarios, operative planning, complication management, and whole-procedure assessment to better approximate real-world neurosurgery.[48,56-60,76-80,116]

Fourth, from a technical standpoint, the current instrument tracking system using optical fiducials requires physical marker setup and introduces practical constraints. While optical tracking with fiducial markers provides the high-precision, real-time three-dimensional kinematic data required

for accurate EV-ICEMS metrics, optimizing this system for the intelligent operating room will require lighter tracking markers, expanded camera coverage, ideally upgraded to four cameras to prevent tracking data loss, and lighter, lower-profile fiducial markers that reduce instrument handling interference and marker-related occlusion in the surgical field—for example, during periods of active bleeding or instrument crowding.[34] Computer vision-based tracking, while appealing for its marker-free convenience, faces significant challenges in the surgical environment, including occlusion from blood, instruments in field of vision, lighting variability, and limited depth accuracy for precise tool-tip kinematics. Deformation and viscoelastic feature modeling of soft tissue adds additional computational complexity that current computer vision pipelines do not robustly address.[44,90-93]

Fifth, the current study is limited in sample size and institutional scope. Although the pre-task questionnaire included broad surgical interest items for medical students and self-reported prior technical experience related to subpial resection in epilepsy and tumor cases, these variables were not uniformly collected across all participants. These factors could influence task familiarity, motivation, and performance, offering a possible explanation for composite score variability among the heterogeneous expertise groups. Future studies should standardize these items by recording recent and lifetime subpial corticectomy volume, subspecialty exposure, operating room exposure, and learner motivation. Prospective multi-institutional trials with larger participant samples across multiple training programs will be essential to externally validate the EV-ICEMS model and confirm their generalizability.

Sixth, while the ex vivo calf brain subpial corticectomy model is the focus of this work, the EV-ICEMS framework is in principle applicable to any surgical procedure for which multi-instrument kinematic data can be captured. Within neurosurgery, extending this approach to more complex

bimanual microsurgical procedures, such as temporal lobectomy, represents a natural next step. Many bimanual microsurgical or minimally invasive procedures could benefit from our model. This intelligent tutoring system also carries strong implications for the globalization of safe surgery and global surgical education, extending accessible, equitable, and safety-oriented simulation training to resource-limited and underserved settings.

Future work should also include the emotional and cognitive dimensions of surgical training.[65-71,116] Further, the integration of LLM-based feedback systems capable of generating contextually sensitive, emotionally attuned, and personalized instructions adapted to each learner's behavioral profiles or level of training using EV-ICEMS kinematic data.[34,43,44,114] This may enable the full pedagogical potential of AI-powered surgical training. Transformers being the foundational infrastructure of LLMs, this transition is relatively natural. For learners and educators, EV-ICEMS provides an explainable AI with teachable metrics and instrument-specific transparency needed to build AI-enhanced surgical curricula and next-generation intelligent tutoring systems.[25,28,31,32,54] The complementary finding from our group's recent RCT that AI-augmented personalized expert instruction underscores that the future of intelligent surgical education lies in human-in-the-loop AI-educator/learner synergy.[27,30,105]

Beyond technical instrumentation, future iterations of the EV-ICEMS must deliberately simulate the full complexity of the operative environment, including the integration of its failures.[28–30,34,56-60] Controlled surgical mistakes and failure scenarios such as an actively bleeding vessel, unexpected anatomical variation, or instrument malfunction would allow trainees to practice high-stakes risk scenarios and error recognition in a consequence-free setting, building the adaptive resilience and OR preparedness that real operative crises demand.[56-60]

Finally, ethical considerations must guide the translation of EV-ICEMS to clinical and educational practice. The deployment of AI systems in high-stakes domains such as surgical training requires robust AI safety frameworks, transparency, and protection against algorithmic bias.[48,91,92,110,115,116] Responsible AI translation to the intelligent OR is a prerequisite for earning the trust of surgeons, trainees, educators, and patients alike.[48,91,92,110,115,116] Ultimately, every technical and pedagogical development described in this thesis must be anchored to its terminal purpose: patient safety and agency in patient-centered care. Patient-centered agency, i.e., empowering the patient as an active stakeholder whose stories, challenges, and outcomes must guide surgical decision-making, should therefore be embedded as a foundational principle in the ethical translation of EV-ICEMS from simulation to clinical practice. Our long-term vision is an AI-powered and human-supervised intelligent OR in which real-time kinematic assessment, LLM-generated personalized feedback, and expert human mentorship operate in concert to maximize learning outcomes while preserving patient safety.

Conclusion

This study outlines the development and validation of EV-ICEMS, a deep learning methodology to differentiate surgical expertise levels, enabling the optimization of intelligent tutoring systems. The Transformer architecture trained on data from a calf brain subpial corticectomy model demonstrated strong construct validity and superior predictive validity for simulation training of bimanual procedural interventions in realistic surgical settings. Pairwise tool-metric analyses identified microscissors and bipolar kinematic metrics as the most important predictors of surgical expertise, consistent with the procedural demands of the subpial corticectomy task.

REFERENCES

1. Canadian Institute for Health Information. *Hospital Harm.* CIHI; 2025. https://www.cihi.ca/en/indicators/hospital-harm

2. Dao Trong P, Olivares A, El Damaty A, Unterberg A. Adverse events in neurosurgery: a comprehensive single-center analysis of a prospectively compiled database. *Acta Neurochir.* 2023;165(3):585-593. doi:10.1007/s00701-022-05462-w

3. Stone S, Bernstein M. Prospective Error Recording in Surgery: An Analysis of 1108 Elective Neurosurgical Cases. *Neurosurgery.* 2007;60(6):1075-1082. doi:10.1227/01.NEU.0000255466.22387.15

4. Rogers SO, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery.* 2006;140(1):25-33. doi:10.1016/j.surg.2006.01.008

5. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. *Ann Surg.* 2017;265(3):492-501. doi:10.1097/SLA.0000000000001959

6. Kotsis S, Chung K. Application of the "See One, Do One, Teach One?" Concept in Surgical Training. *Plast Reconstr Surg.* 2013;131(5):1194-1201. doi:10.1097/PRS.0b013e318287a0b3

7. Camison L, Brooker JE, Naran S, Potts JRI, Losee JE. The History of Surgical Education in the United States: Past, Present, and Future. *Annals of Surgery Open.* 2022;3(1):e148. doi:10.1097/as9.0000000000000148

8. Polavarapu HV, Kulaylat AN, Sun S, Hamed OH. 100 years of surgical education: the past, present, and future. *Bull Am Coll Surg.* 2013;98(7):22-7.

9. Cianciolo AT, Blessman J. “See One, Do One, Teach One?”A Story of How Surgeons Learn. In: Kohler TS, Schwartz B, eds. *Surgeons as Educators.* Springer, Cham; 2017:3-13. doi: 10.1007/978-3-319-64728-9_1

10. Mirza M, Koenig JF. Teaching in the Operating Room. In: Kohler TS, Schwartz B, eds. *Surgeons as Educators*. Springer, Cham; 2017:137-160. doi:10.1007/978-3-319-64728-9_8

11. Rowse PG, Dearani JA. Deliberate Practice and the Emerging Roles of Simulation in Thoracic Surgery. *Thorac Surg Clin.* 2019;29(3):303-309. doi:10.1016/j.thorsurg.2019.03.007

12. Roberts NK, Williams RG, Kim MJ, Dunnington GL. The Briefing, Intraoperative Teaching, Debriefing Model for Teaching in the Operating Room. *J Am Coll Surg.* 2009;208(2):299-303. doi:10.1016/j.jamcollsurg.2008.10.024

13. Ericsson KA. Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Acad Med*. 2015;90(11):1471-1486. doi:10.1097/ACM.0000000000000939

14. Kitto S, Fantaye AW, Zevin B, et al. A Scoping Review of the Literature on Entrustable Professional Activities in Surgery Residency Programs. *J Surg Educ*. 2024;81(6):823-840. doi:10.1016/j.jsurg.2024.02.011

15. Royal College of Physicians and Surgeons of Canada. *Competence by Design.* https://www.royalcollege.ca/ca/en/cbd/understanding-cbd/the-rationale-for-change.html.

16. Frank JR, Snell LS, Cate O Ten, et al. Competency-based medical education: theory to practice. *Med Teach.* 2010;32(8):638-645. doi:10.3109/0142159X.2010.501190

17. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x

18. Wagner JP, Lewis CE, Tillou A, et al. Use of Entrustable Professional Activities in the Assessment of Surgical Resident Competency. *JAMA Surg.* 2018;153(4):335-343. doi:10.1001/jamasurg.2017.4547

19. Hackney L, O'Neill S, O'Donnell M, Spence R. A scoping review of assessment methods of competence of general surgical trainees. *Surgeon.* 2023;21(1):60-69. doi:10.1016/j.surge.2022.01.009

20. van Hove PD, Tuijthof GJ, Verdaasdonk EG, et al. Objective assessment of technical surgical skills. *Br J Surg.* 2010;97(7):972-87. doi:10.1002/bjs.7115

21. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med*. 1996;71(12):1363-5. doi:10.1097/00001888-199612000-00023

22. Orovec A, Bishop A, Scott SA, et al. Validation of a Surgical Objective Structured Clinical Examination (S-OSCE). *J Surg Educ.* 2022;79(4):1000-1008. doi:10.1016/j.jsurg.2022.01.014

23. Alotaibi FE, AlZhrani GA, Sabbagh AJ, et al. Neurosurgical Assessment of Metrics Including Judgment and Dexterity (NAJD Metrics). *Surg Innov.* 2015;22(6):636-42. doi:10.1177/1553350615579729

24. Tobin S. Entrustable Professional Activities in Surgical Education. In: Nestel D, et al., eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Springer, Singapore. 2019; 17:229-238. doi:10.1007/978-981-13-3128-2_21

25. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One.* 2020;15(2):e0229596. doi:10.1371/journal.pone.0229596

26. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw Open.* 2022;5(2):e2149008. doi:10.1001/jamanetworkopen.2021.49008

27. Giglio B, Albeloushi A, Alhaj AK, et al. Artificial intelligence-augmented human instruction and surgical simulation performance: a randomized clinical trial. *JAMA Surg.* 2025;160(9):993-1003. doi:10.1001/jamasurg.2025.2564

28. Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *npj Digit Med.* 2022;5(1):54. doi:10.1038/s41746-022-00596-8

29. Yilmaz R, Bakhaidar M, Alsayegh A, et al. Real-time multifaceted artificial intelligence vs in-person instruction in teaching surgical technical skills: a randomized controlled trial. *Sci Rep*. 2024;14(1):15130. doi:10.1038/s41598-024-65716-8

30. Yilmaz R, Alsayegh A, Bakhaidar M, et al. Combining real-time AI and in-person expert instruction in simulated surgical skills training - randomized crossover trial. *npj Artif Intell.* 2025;1(1):36. doi:10.1038/s44387-025-00032-8

31. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open*. 2019;2(8):e198363. doi:10.1001/jamanetworkopen.2019.8363

32. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task. *J Bone Joint Surg Am*. 2019;101(23):e127. doi:10.2106/JBJS.18.01197

33. Alsayegh A, Bakhaidar M, Winkler-Schwartz A, Yilmaz R, Del Maestro RF. Best practices using ex vivo animal brain models in neurosurgical education to assess surgical expertise. *World Neurosurg*. 2021;155:e369-e381. doi:10.1016/j.wneu.2021.08.061

34. Almansouri A, Abou Hamdan N, Yilmaz R, et al. Continuous instrument tracking in a cerebral corticectomy ex vivo calf brain simulation model: face and content validation. *Oper Neurosurg* (Hagerstown). 2024;27(1):106-113. doi:10.1227/ons.0000000000001044

35. Winkler-Schwartz A, Yilmaz R, Tran DH, et al. Creating a comprehensive research platform for surgical technique and operative outcome in primary brain tumor neurosurgery. *World Neurosurg*. 2020;144:e62-e71. doi:10.1016/j.wneu.2020.07.209

36. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arxiv*. 2017; 1-15. doi:10.48550/arXiv.1706.03762

37. Bravo A, Razzaq S, Murari M, Ahuja A, Birchett D, Schumacher L. Transformers in surgical artificial intelligence: a domain-stratified, study-level narrative review. *JTCVS Open*. 2026; 30:101597. doi:10.1016/j.xjon.2026.101597

38. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

39. Lin CH, Kuo CF. A review of transformers in medical research and health care. *Hand Clin.* 2026;42(1):53-63. doi:10.1016/j.hcl.2025.08.007

40. Takahashi S, Sakaguchi Y, Kouno N, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *J Med Syst.* 2024;48(1):84. doi:10.1007/s10916-024-02105-8

41. Ofori E, Dake DK. Explainable artificial intelligence in LSTM transformer models for student performance analysis. *Discov Computing*. 2025;28:313. doi:10.1007/s10791-025-09814-9

42. Alkhamali EA, Allinjawi A, Ashari RB. Combining transformer, convolutional neural network, and long short-term memory architectures: a novel ensemble learning technique that leverages multi-acoustic features for speech emotion recognition in distance education classrooms. *Appl Sci*. 2024;14(12):5050. doi:10.3390/app14125050

43. Yang S, Zhou F, Mayer L, et al. Large-scale self-supervised video foundation model for intelligent surgery. *npj Digit Med*. 2026;9:220. doi:10.1038/s41746-026-02403-0

44. Yilmaz R, Del Maestro RF, Donoho D. Continuous intraoperative AI monitoring of surgical technical skills using computer vision. *Am J Surg*. 2025;245:116248. doi:10.1016/j.amjsurg.2025.116248

45. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc.* 2003;17:1525-1529. doi:10.1007/s00464-003-0035-4

46. Hughes T, Fennelly JT, Patel R, Baxter J. The Validation of Surgical Simulators: A Technical Report on Current Validation Terminology. *Cureus.* 2022;14(11):e31881. doi:10.7759/cureus.31881

47. Borgersen NJ, Naur TMH, Sorensen SMD, et al. Gathering Validity Evidence for Surgical Simulation: A Systematic Review. *Annals of Surgery.* 2018;267(6):1063-1068. doi:10.1097/sla.0000000000002652

48. Marshall SD, Nataraja RM. Patient safety and surgical education. In: Nestel D, Dalrymple K, Paige JT, Aggarwal R, eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Springer, Singapore. 2019;17:327-337. doi:10.1007/978-981-13-3128-2_21

49. Yilmaz R, Ledwos N, Sawaya R, et al. Nondominant hand skills spatial and psychomotor analysis during a complex virtual reality neurosurgical task-a case series study. *Oper Neurosurg* (Hagerstown). 2022;23(1):22-30. doi:10.1227/ons.0000000000000232

50. Asoodar M, Janesarvatan F, Yu H, de Jong N. Theoretical foundations and implications of augmented reality, virtual reality, and mixed reality for immersive learning in health professions education. *Adv Simul* (Lond). 2024;9(1):36. doi:10.1186/s41077-024-00311-5

51. Harley JM, Tawakol T, Azher S, et al. The role of artificial intelligence, performance metrics, and virtual reality in neurosurgical education: an umbrella review. *Glob Surg Educ*. 2024;3:83. doi:10.1007/s44186-024-00284-z

52. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: Best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681-1690. doi:10.1016/j.jsurg.2019.05.015

53. Han JJ, Patrick WL. See one-practice-do one-practice-teach one-practice: importance of practicing outside the operating room. *J Thorac Cardiovasc Surg*. 2019;157(2):671-677. doi:10.1016/j.jtcvs.2018.07.108

54. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. *JAMA Netw Open.* 2023;6(9):e2334658. doi:10.1001/jamanetworkopen.2023.34658

55. Gil-Robles S, Duffau H. Surgical management of WHO Grade II gliomas in eloquent areas: The necessity of preserving a margin around functional structures. *Neurosurg Focus.* 2010;28(2). doi:10.3171/2009.12.FOCUS09236

56. Zigmont JJ, Kappus LJ, Sudikoff SN. Theoretical foundations of learning through simulation. *Semin Perinatol*. 2011;35(2):47-51. doi:10.1053/j.semperi.2011.01.002

57. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. Acad Med. 2003;78(8):783-8. doi:10.1097/00001888-200308000-00006

58. Tan SSY, Sarker SK. Simulation in surgery: a review. *Scott Med J*. 2011;56(2):104-109. doi:10.1258/smj.2011.011098

59. Bjerrum F, Thomsen ASS, Nayahangan LJ, Konge L. Surgical simulation: Current practices and future perspectives for technical skills training. *Medical Teacher*. 2018;40(7):668-675. doi:10.1080/0142159X.2018.1472754

60. de Montbrun SL, Macrae H. Simulation in surgical education. *Clin Colon Rectal Surg*. 2012;25(3):156-65. doi:10.1055/s-0032-1322553

61. Aggarwal R. Simulation in Surgical Education. In: Nestel D, et al., eds. *Advancing Surgical Education*. Springer, Singapore. 2019;17:269-278. doi:10.1007/978-981-13-3128-2_21

62. Badash I, Burtt K, Solorzano CA, Carey JN. Innovations in surgery simulation: a review of past, current and future techniques. *Ann Transl Med*. 2016;4(23):453. doi:10.21037/atm.2016.12.24

63. Cao C, Cerfolio RJ. Virtual or Augmented Reality to Enhance Surgical Education and Surgical Planning. *Thorac Surg Clin*. 2019;29(3):329-337. doi:10.1016/j.thorsurg.2019.03.010

64. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Oper Neurosurg* (Hagerstown). 2012;71:32-42. doi:10.1227/NEU.0b013e318249c744

65. Duffy MC, Lajoie SP, Pekrun R, Lachapelle K. Emotions in medical education: Examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. *Learn Instr*. 2020;70:101150. doi:10.1016/j.learninstruc.2018.07.001

66. Chandler P, Sweller J. Cognitive Load Theory and the format of instruction. *Cogn Instr*. 1991;8(4):293-332. doi:10.1207/s1532690xci0804_2

67. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: Implications for medical education: AMEE Guide No. 86. *Med Teach*. 2014;36(5):371-384. doi:10.3109/0142159X.2014.889290

68. Artino AR, Holmboe ES, Durning SJ. Can achievement emotions be used to better understand motivation, learning, and performance in medical education? *Med Teach*. 2012;34(3):240-244. doi:10.3109/0142159X.2012.643265

69. Pekrun R. The Control-Value Theory of Achievement Emotions. *Educ Psychol Rev*. 2006;18:315-341. doi:10.1007/s10648-006-9029-9

70. Engel GL. The need for a new medical model: a challenge for biomedicine. *Science*. 1977;196(4286):129-136. doi:10.1126/science.847460

71. Tze V, Parker P, Sukovieff A. Control-Value Theory of Achievement Emotions and Its Relevance to School Psychology. *Can J Sch Psychol*. 2022;37(1):23-39. doi:10.1177/08295735211053962

72. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg*. 2002;236(4):458-464. doi:10.1097/00000658-200210000-00008

73. Yudkowsky R, Luciano C, Banerjee P, et al. Practice on an augmented reality/haptic simulator and library of virtual brains improves residents ability to perform a ventriculostomy. *Simul Healthc*. 2013;8(1):25-31. doi:10.1097/SIH.0b013e3182662c69

74. Thawani JP, Ramayya AG, Abdullah KG, et al. Resident simulation training in endoscopic endonasal surgery utilizing haptic feedback technology. *J Clin Neurosci*. 2016;34:112-116. doi:10.1016/j.jocn.2016.05.036

75. Tran DH, Winkler-Schwartz A, Tuznik M, et al. Quantitation of Tissue Resection Using a Brain Tumor Model and 7-T Magnetic Resonance Imaging Technology. *World Neurosurg*. 2021;148(4):e326-e339. doi:10.1016/j.wneu.2020.12.141

76. Kirkman MA, Ahmed M, Albert AF, et al. The use of simulation in neurosurgical education and training: A systematic review. *J Neurosurg*. 2014;121(2):228-246. doi:10.3171/2014.5.JNS131766

77. Chan J, Pangal DJ, Cardinal T, et al. A systematic review of virtual reality for the assessment of technical skills in neurosurgery. *Neurosurg Focus*. 2021;51(2):E15. doi:10.3171/2021.5.FOCUS21210

78. Davids J, Manivannan S, Darzi A, et al. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. *Neurosurg Rev.* 2021;44(4):1853-1867. doi:10.1007/s10143-020-01378-0

79. Oliveira L, Figueiredo E. Simulation training methods in neurological surgery. Asian J *Neurosurg*. 2019;14(02):364-370. doi:10.4103/ajns.AJNS_269_18

80. Limbrick DD, Dacey RG. Simulation in Neurosurgery. *Neurosurgery.* 2013;73(Suppl 1):S1-S3. doi:10.1227/NEU.0000000000000094

81. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. *Oper Neurosurg.* 2015;11(1):89-98. doi:10.1227/NEU.0000000000000631

82. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *Int J Comput Assist Radiol Surg*. 2015;10(5):603-618. doi:10.1007/s11548-014-1091-z

83. Penfield W, Jasper HH. *Epilepsy and the Functional Anatomy of the Human Brain.* 1st ed. Little Brown; 1954.

84. Al-Otaibi F, Baeesa SS, Parrent AG, Girvin JP, Steven D. Surgical techniques for the treatment of temporal lobe epilepsy. *Epilepsy Res Treat*. 2012;2012:374848. doi:10.1155/2012/374848

85. Silbergeld D, Hebb A, Yang T. The sub-pial resection technique for intrinsic tumor surgery. *Surg Neurol Int.* 2011;2(1):180. doi:10.4103/2152-7806.90714

86. Sabbagh AJ, Bajunaid KM, Alarifi N, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurg*. 2020;139:e220-e229. doi:10.1016/j.wneu.2020.03.187

87. Hicdonmez T, Hamamcioglu MK, Tiryaki M, et al. Microneurosurgical training model in fresh cadaveric cow brain: laboratory study simulating approach to the circle of Willis. *Surg Neurol*. 2006;66(1):100-104. doi:10.1016/j.surneu.2005.09.027

88. Carey JN, Minneti M, Leland HA, et al. Perfused fresh cadavers: method for application to surgical simulation. *Am J Surg*. 2015;210(1):179-187. doi:10.1016/j.amjsurg.2014.10.027

89. Schmidt MJ, Pilatus U, Wigger A, et al. Neuroanatomy of the calf brain as revealed by high-resolution magnetic resonance imaging. *J Morphol*. 2009;270(6):745-758. doi:10.1002/jmor.10717

90. Sahhar M, Nassar JE, Rouffiac AE, et al. Enhanced surgical instructions using large language models. *JAMA Surg*. 2025;160(10):1158-1159. doi:10.1001/jamasurg.2025.2825

91. Eke CI, Shuib L. The role of explainability and transparency in fostering trust in AI healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Comput Appl*. 2024. doi:10.1007/s00521-024-10868-x

92. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable AI in healthcare: A systematic review. *Heliyon*. 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110

93. Rengers TA, Thiels CA, Salehinejad H. Academic surgery in the era of large language models: a review. *JAMA Surg.* 2024;159(4):445-450. doi:10.1001/jamasurg.2023.6496

94. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract.* 2014;19(2):233-250. doi:10.1007/s10459-013-9458-4.

95. Schout BM, Hendrikx AJ, Scheele F, et al. Validation and implementation of surgical simulators: a critical review of present, past, and future. *Surg Endosc*. 2010;24(3):536-546. doi:10.1007/s00464-009-0634-9

96. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: Promises and perils. *Ann Surg.* 2018;268(1):70-76. doi:10.1097/SLA.0000000000002693

97. Liao W, Zhu Y, Zhang H, et al. Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. *Front Surg*. 2025;12:1551838. doi:10.3389/fsurg.2025.1551838

98. Leon S, Lee S, Perez JE, Hashimoto DA. Artificial intelligence and the education of future surgeons. *Am J Surg*. 2025;246:116257. doi:10.1016/j.amjsurg.2025.116257

99. Guni A, Varma P, Zhang J, Fehervari M, Ashrafian H. Artificial intelligence in surgery: the future is now. *Eur Surg Res*. 2024;65(1):22-39. doi:10.1159/000536393

100. Vannaprathip N, Haddaway P, Schultheis H, Suebnukarn S. Intelligent Tutoring for Surgical Decision Making: a Planning-Based Approach. *Int J Artif Intell Educ*. 2022;32:350-381. doi:10.1007/s40593-021-00261-3

101. Mousavinasab E, Zarifsanaiey N, et al. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interact Learn Environ.* 2021;29(1):142-163. doi:10.1080/10494820.2018.1558257

102. Julian D, Smith R. Developing an intelligent tutoring system for robotic-assisted surgery instruction. *Int J Med Robot*. 2019;15(6):e2037. doi:10.1002/rcs.2037

103. Furlan R, Gatti M, Mene R, et al. A Natural Language Processing-Based Virtual Patient Simulator and Intelligent Tutoring System. *JMIR Med Inform*. 2021;9(4):e24073. doi:10.2196/24073

104. Mirchi N, Ledwos N, Del Maestro RF. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. *Can J Neurol Sci*. 2021;48(2):198-200. doi:10.1017/cjn.2020.202

105. Davidovic V, Giglio B, Albeloushi A, et al. Effect of artificial intelligence-augmented human instruction on feedback frequency and surgical performance during simulation training. *J Surg Educ*. 2025;82(11):103743. doi:10.1016/j.jsurg.2025.103743

106. Yilmaz R, Fazlollahi A, Alsayegh A, et al. 428 Artificial Intelligence Training Versus In-person Expert Training - Cross-Over RCT. *Neurosurgery*. 2024;70(Suppl 1):129-130. doi:10.1227/neu.0000000000002809_428

107. Yilmaz R, Bakhaidar M, Alsayegh A, et al. Comparing the Efficiency of a Real-Time Artificial Intelligence Instructor to Human Expert Instructors in Simulated Surgical Technical Skills Training. *Neurooncol Adv*. 2023;5(Suppl 2):i1-i1. doi:10.1093/noajnl/vdad071.003

108. Godbole GH, Hawkins D, Ewool K, Camacho MH, Karim R, Patel B. Educational impact of automated feedback systems in surgical training: a systematic review with quantitative synthesis. *J Surg Educ*. 2026;83(4):103879. doi:10.1016/j.jsurg.2026.103879

109. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. doi:10.1016/j.jbi.2020.103655

110. Lang BH, Nyholm S, Blumenthal-Barby J. Responsibility Gaps and Black Box Healthcare AI: Shared Responsibilization as a Solution. *Digit Soc*. 2023;2:52. doi:10.1007/s44206-023-00073-z

111. Netland T, von Dzengelevski O, Tesch K, Kwasnitschka D. Comparing human-made and AI-generated teaching videos: An experimental study on learning effects. *Comput Educ.* 2025;224:105164. doi:10.1016/j.compedu.2024.105164

112. Pogue LL, AhYun K. The Effect of Teacher Nonverbal Immediacy and Credibility on Student Motivation and Affective Learning. *Commun Educ*. 2006;55(3):331-344. doi:10.1080/03634520600748623

113. Siegle RF, Craig SD. The voice quality of pedagogical agents impacts learning and agent perceptions. *J Comput Assist Learn*. 2024;40:2278-2291. doi:10.1111/jcal.13027

114. Pugh CM, Alikhani MA, Rodman A. Intelligent, human-centric delivery is needed to maximize AI. *JAMA Surg*. 2025;160(9):1003. doi:10.1001/jamasurg.2025.2550

115. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet. Digital health*. 2020; *2*(10), e537–e548. doi:10.1016/S2589-7500(20)30218-1

116. Miyasaka K. Approaching surgery simulation education from a patient-centered pathway. In: Nestel D, Dalrymple K, Paige JT, Aggarwal R, eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Springer, Singapore. 2019;17:451-456.

117. Zhang H, Wang J, Liu C, et al. Development of a continuously perfused ex vivo kidney training model for laparoscopic partial nephrectomy: validity and efficiency. *Int J Surg.* 2023;109(12):3919-3928. doi:10.1097/JS9.0000000000000753

118. Filiz M, Yilmaz R, Tee T, et al. AI assessment of surgical technical skill adaptation across depth levels in simulated tumor resection: a case series study. *J. Neurosurg*. 1, 1-11. Published online May 1, 2026. doi: 10.3171/2025.12.JNS251528

119. R Core Team. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing. Accessed January 28, 2026. https://www.R-project.org/

120. Skjold-Odegaard B, Soreide K. Competency-based Surgical Training and Entrusted Professional Activities - Perfect Match or a Procrustean Bed? *Annals of Surgery*. 2021;273(5):e173-e175. doi:10.1097/sla.0000000000004521

TABLES

Table 1. Validity Frameworks in Surgical Simulation

Framework	Approach	Components
Traditional	Operationalize distinct types of validity as discrete attributes of a measurement tool.	**Face**: Whether the simulation task appears, to expert users, to resemble the actual operative procedure it is designed to represent. **Content**: the extent to which the simulator trains the skills it is created to train. **Construct**: the ability of the simulator (or of a scoring system used with it) to accurately differentiate between individuals of different skill levels (expertise groups). **Predictive**: Whether scores (or metrics) derived from simulator performance generalize to, and anticipate, a participant's future performance in tasks or operative settings from previously unseen data
Messick's Contemporary	Create a cumulative, multi-source validity argument in which each piece of evidence strengthens the case for the intended interpretation and use of model.	**Test content**: The degree to which the tasks and items within the assessment instrument are representative of and relevant to the underlying skill or construct being evaluated. **Response process**: the integrity of data collection, such as standardized instructions and blinded evaluations. **Internal structure**: the degree to which components of an instrument match the underlying construct, as captured by statistical measures (e.g., internal consistency, reliability). **Relationship to other variables**: the extent to which assessment measures are related to external measures (e.g., level of training, experience). **Consequences**: Evidence regarding the intended and unintended outcomes (e.g., educational, ethical, psychosocial, and systemic factors) that result from the deployment of the training system.

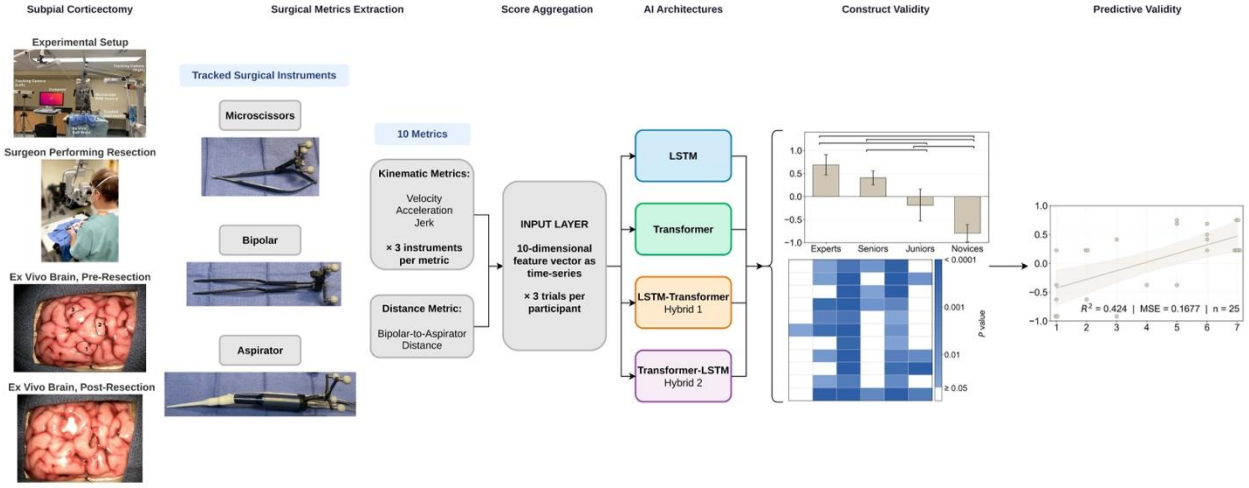
Table 2. Demographic Characteristics of Included Study Participants

Characteristic	No. (%)			
	Novices (n = 14)	**Juniors (n = 14)**	**Seniors (n = 11)**	**Experts (n = 8)**
Age, mean (SD), y	20.4 (2.8)	28.1 (1.9)	35.2 (3.0)	46.1 (8.0)
Sex				
Male	4 (29)	9 (64)	8 (73)	7 (88)
Female	10 (71)	5 (36)	3 (27)	1 (12)
Handedness				
Right	13 (93)	12 (86)	8 (73)	8 (100)
Left	1 (7)	2 (14)	2 (18)	0
Ambidextrous	0	0	1 (9)	0
Prior NeuroVR exposure				
Yes	2 (14)	4 (29)	4 (36)	7 (88)
No	12 (86)	10 (71)	7 (64)	1 (12)
Medical students	14	–	–	–
Residents	–	14	4	–
PGY-1	–	5	–	–
PGY-2	–	3	–	–
PGY-3	–	2	–	–
PGY-4	–	1	–	–
PGY-5	–	3	–	–
PGY-6	–	–	4	–
Fellows	–	–	7	–
Spine	–	–	2	–
Epilepsy	–	–	1	–
Pediatrics	–	–	2	–
Oncology	–	–	1	–
Functional	–	–	1	–
Neurosurgeons	–	–	–	8
Pediatric	–	–	–	3
Oncology	–	–	–	2
Cerebrovascular	–	–	–	1
Skull base	–	–	–	1
Spine/general	–	–	–	1
Primary institution				
McGill University	10	12	11	8
Université de Montréal	4	0	0	0
Université de Sherbrooke	0	2	0	0

Abbreviations: PGY, post-graduate year.

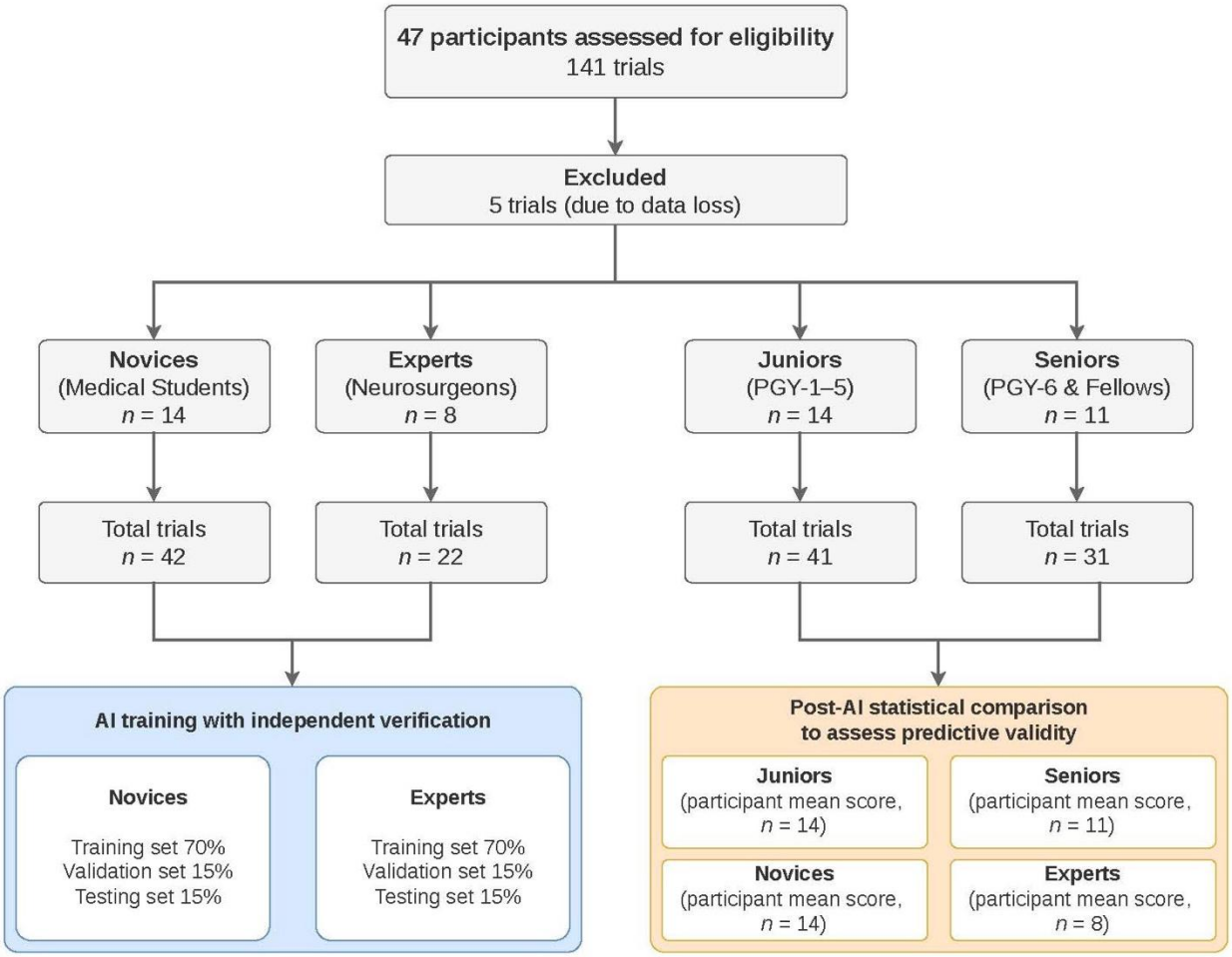
FIGURES

Figure 1. Study Workflow



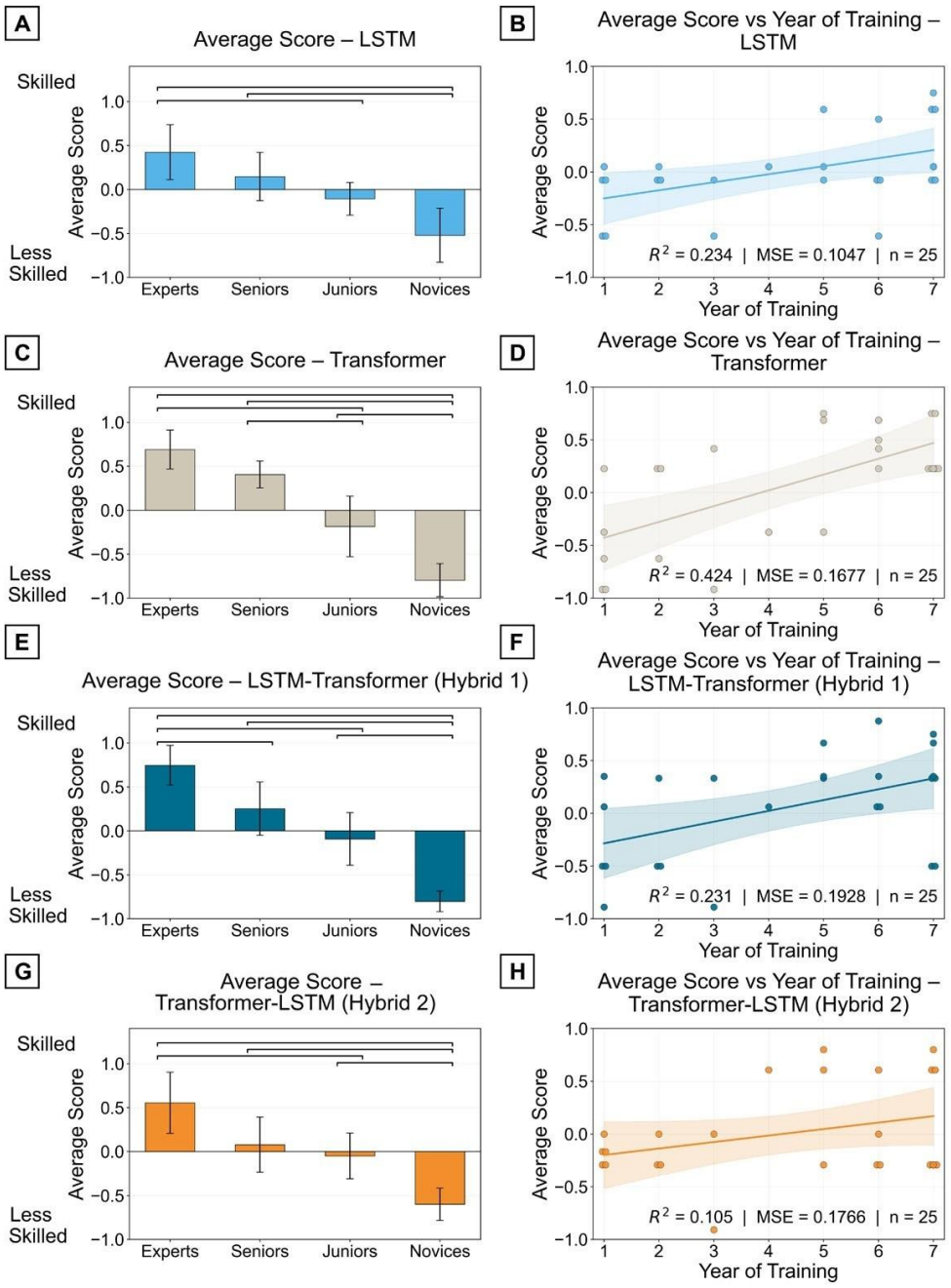
Schematic overview of a deep learning methodology to differentiate surgical expertise levels utilizing metrics extraction, AI model development, and validation. Participants complete 3 subpial resections using an ex vivo calf brain corticectomy model, enabling a realistic and reproducible evaluation of surgical performance. Activity of 3 surgical instruments (microscissors, bipolar, and aspirator) were recorded using the tracking system. For each instrument, three kinematic metrics (velocity, acceleration, and jerk) and inter-instrument distance (bipolar-to-aspirator) were extracted as time-series features. The input data were fed into 4 AI architectures: an LSTM, a Transformer, an LSTM-Transformer (Hybrid 1), and a Transformer-LSTM (Hybrid 2). Models were trained on experts (neurosurgeons) and novices (medical students) and evaluated on all levels of training (experts, seniors, juniors, and novices). Model outputs consisted of a regression-based composite score derived from the average of all tool-metric time-series predicted scores, scaled from –1 (novice) to 1 (expert). Construct validity was assessed by comparing composite scores across training levels. Statistical significance of individual tool-metric pairs and the composite score is shown in the shaded matrix, where white indicates non-significant results ($P \geq .05$), and blue indicates significance ($P < .05$). Predictive validity was evaluated using previously unseen data from intermediate groups (juniors to seniors). Best-fit regression lines and associated metrics quantified the AI models' ability to predict surgical skill across the year of training.

Figure 2. Participant Recruitment Flowchart



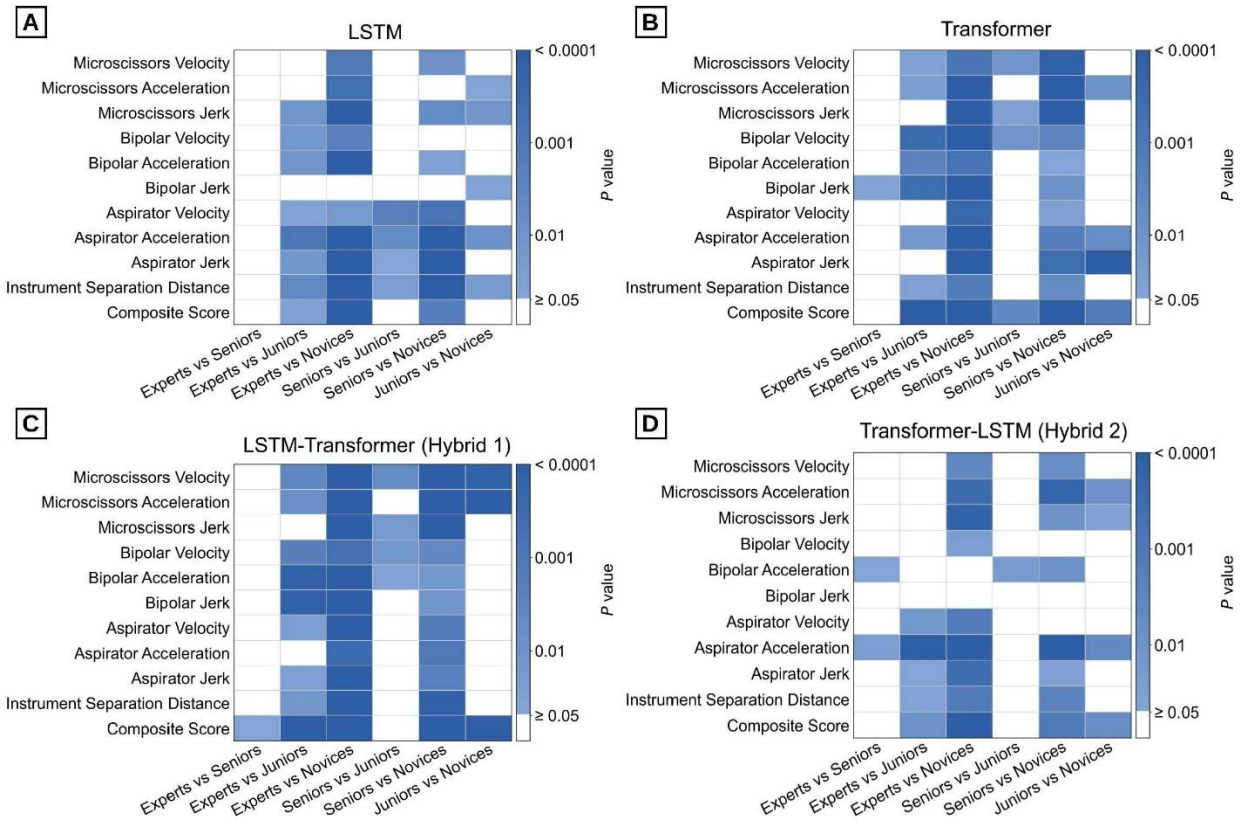
AI indicates artificial intelligence; PGY indicates post-graduate year.

Figure 3. Construct and Predictive Validity of Scores Across Model Architectures



Average scores by training level showing progressive improvement ($P < .05$) with 95% confidence intervals (CI) as error bars and Tukey post hoc significance indicated by horizontal brackets. Panels show participant-level composite scores and regression-based predictive trends for LSTM (A-B), Transformer (C-D), LSTM-Transformer (Hybrid 1) (E-F), and Transformer-LSTM (Hybrid 2) (G-H). The Transformer and Hybrid 1 architectures demonstrated the strongest construct validity, while the Transformer model showed superior predictive validity across training years. The dots represent the average score of each participant, and the x-axis represents year of training. Resident participants' neurosurgery training program was six years. Neurosurgical fellows were considered in seventh year of training. LSTM indicates long short-term memory; MSE indicates mean squared error.

Figure 4. Shaded Matrix Showing Statistical Significance Across Tool-Metric Pairs and Composite Scores for 4 Model Architectures



White indicates non-significant comparisons, while blue indicates statistical significance ($P < .05$). (A) LSTM, (B) Transformer, (C) LSTM-Transformer (Hybrid 1), and (D) Transformer-LSTM (Hybrid 2). LSTM indicates long short-term memory.

APPENDIX A– The Study Supplementary Information

eAppendix A1. Model Design, Architecture, and Hyperparameters

The Intelligent Continuous Expertise Monitoring System (ICEMS) validated in the virtual reality (VR) setting uses a basic Long Short-Term Memory (LSTM) network architecture for time-series prediction and pattern analysis. The LSTM architecture, as a subset of recurrent neural networks (RNN), excels at preserving long-term dependencies in sequential data through gated memory cells, making it well-suited for local temporal pattern discovery.[1] Given the limited complex, noisy, and real-time dynamics of the ex vivo model data captured, we implemented Transformer

and Transformer-based hybrid architectures (i.e., LSTM-Transformer [Hybrid 1] and Transformer-LSTM [Hybrid 2]) to optimize expertise-based score prediction as intelligent tutoring systems. The Transformer's attention mechanism enables predictive modeling of global contextual relationships, capturing nonlinear, adaptive, and dynamic interactions across the full temporal sequence.[2] All 4 AI model architectures processed time-series windows of stacked instrument-tools metrics per participant.

The LSTM model consisted of a single recurrent input layer (LSTM, 128 units) followed by a Dropout layer of 0.30, batch normalization, a fully connected Dense layer (32 units, ReLU), another Dropout 0.30 layer, and a final fully connected Dense layer (1, tanh output) for the regression-based scoring (eFigure 1). The 30% dropout mitigates overfitting. We used a single LSTM layer (128 units) across models to ensure a standardized comparison across the 4 AI architectures. This also reduced the risk of overfitting and was found most suitable for our dataset size. Experimentation with multiple stacked LSTM layers decreased performance, supporting the use of a single layer architecture.

The Transformer model consisted of an input projection (Dense 64), learnable positional embeddings, and 2 stacked encoder blocks with multi-head self-attention (4 heads, key_dim = 32), followed by feed-forward layers with residual connections and layer normalization. Dropout (0.30) was applied after attention layers and within feed-forward layers of each encoder block, as well as after the dense projection layers. The architecture concluded with global average pooling, a fully connected Dense layer (32 units, ReLU), Dropout (0.30), and a final Dense layer (1, tanh) as the output regression layer (**eFigure 1**).

The LSTM-Transformer (Hybrid 1) was designed to first model sequential dependencies using the LSTM's recurrent architecture and subsequently integrate cross-metric relationships and long-range temporal dependencies using the transformer's attention layers. Its architecture consisted of the following components in order from the input layer: LSTM (128 units, return_sequences = True), followed by Dropout 0.30, Dense 64 layer, positional embedding, Transformer encoder block with Multi-Head Attention (4 heads, key_dim = 32), feed-forward layers with residual connections and layer normalization, and dropout applied after attention and within feed-forward layers, GlobalAveragePooling1D, fully connected Dense 32 (ReLU), Dropout 0.30, and a fully connected Dense 1 (tanh) as output regression layer.

Conversely, the Transformer-LSTM hybrid (Hybrid 2) reverses this ordering, applying attention-based global feature integration prior to the LSTM's sequential temporal modeling.[1-4] The algorithm's architecture consists of an input projection (Dense 64), learnable positional embedding, and a Transformer encoder block with Multi-Head Attention (4 heads, key_dim = 32), feed-forward layers with residual connections, layer normalization, and dropout 0.30 applied after attention and within feed-forward layers. This Transformer unit was followed by an LSTM layer (128 units), batch normalization, a fully connected Dense (32 units, ReLU), a dropout 0.30 layer, and lastly a Dense layer (1, tanh output) for the regression output.

All models were implemented in Python (version 3.12.3) using TensorFlow (version 2.19.0) and trained under identical hyperparameter settings (Adam optimizer, learning rate [lr] 1×10^{-3}, mean-squared error [MSE] loss, 50 epochs, batch size 32, and early stopping with patience = 8 based

on validation loss), ensuring differences in performance reflected architectural variation rather than training configuration. Data were split 70/15/15 into train/validation/test sets, stratified by expertise group and participant to prevent leakage across trials. The 30% dropout and batch normalization improved regularization and generalizability. Models were trained exclusively on expert and novice data. Predictions were generated at the trial level, averaged per participant to obtain composite scores.

Supplemental References

1. Alkhamali EA, Allinjawi A, Ashari RB. Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms. *Appl. Sci.* 2024;14(12):5050. doi:10.3390/app14125050
2. Winkler-Schwartz A, Yilmaz R, Tran DH, et al. Creating a comprehensive research platform for surgical technique and operative outcome in primary brain tumor neurosurgery. *World Neurosurg*. 2020;144:e62-e71. doi:10.1016/j.wneu.2020.07.209
3. Bravo A, Razzaq S, Murari M, Ahuja A, Birchett D, Schumacher L. Transformers in Surgical Artificial Intelligence: A Domain-Stratified, Study-Level Narrative Review. *JTCVS Open*. 2026:101597.
4. Ofori E, Dake DK. Explainable artificial intelligence in LSTM transformer models for student performance analysis. *Discov Computing*. 2025;28:313. doi:10.1007/s10791-025-09814-9
5. Delorme S, Laroche D, DiRaddo R, & Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Operative Neurosurg*. 2012;71:32–42.

doi:10.1227/NEU.0b013e318249c744

6. Liao W, Zhu Y, Zhang H, et al. Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. *Front Surg*. 2025;12:1551838. doi:10.3389/fsurg.2025.1551838

eAppendix A2. Composite Score Derivation

Scores were computed exclusively between the novice (–1) and expert (1) groups for each instrument–metric pair using each of the four AI model architectures.

Four interpretable and coachable metrics were derived for each of the three tools (microscissors, bipolar and aspirator): 3 kinematic features (velocity, acceleration, jerk) and inter-instrument separation distance (bipolar-to-aspirator). This resulted in a total of 10 unique tool-metric pairs per participant per trial. Each of the 4 AI models was trained exclusively on the expert and novice group time-series data for each of these tool-metric pairs. For each participant and trial, a regression score was obtained for all ten tool-metric pairs. The final composite score was obtained by averaging these regression scores across all tool-metric pairs and across the three trials per participant. Due to some data loss from the instrument tracking system, 5 of the 47 participants had only two trials available instead of three. The details of these five participants are provided in **eTable 1**.

eTable 1. Missing Participant Trials

Participant ID	Missing Trial	Expertise Group	Level of Training
10290713	Trial 3	Junior	PGY-1
16041012	Trial 2	Senior	PGY-6
32030213	Trial 1	Senior	Functional Fellow
46051812	Trial 1	Expert	Spine Neurosurgeon (Staff)
25161110	Trial 3	Expert	Pediatric Neurosurgeon (Staff)

Among N = 47 participants each performing 3 trials, we initially had data from 141 trials. However, data from 5 trials were lost due to technical issues with the tracking system. Thus, a total of 136 trials were used for model training and assessment. The 5 lost trials included 1 junior, 2 senior, and 2 expert participants. PGY, post-graduate year.

eTable 2. Metric Extraction from Real-Time Tracking Data

Metric	Description	Formula
Instrument Position (IP)	3D coordinates of the tool tip at time t (mm)	$X(t), Y(t), Z(t)$
Velocity (V) (mm/s)	Instantaneous 3D speed between successive samples	$V(t) = \text{sqrt}\,[(X(t) - X(t-1))^2 + (Y(t) - Y(t-1))^2 + (Z(t) - Z(t-1))^2] / (T(t) - T(t-1))$
Acceleration (A) (mm/s^2)	Rate of change of velocity	$A(t) = (V(t) - V(t-1)) / (T(t) - T(t-1))$
Jerk (J) (mm/s^3)	Rate of change of acceleration	$J(t) = (A(t) - A(t-1)) / (T(t) - T(t-1))$
Inter-Instrument Distance (IID) (mm)	Moment-to-moment 3D spacing between instrument tool tips (bipolar-to-aspirator)	$IID\,(t) = \text{sqrt}\,[(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2]$

Derived kinematic performance metrics were computed from sequential instrument tracking data (10 Hz sampling, change in time [Δt] = 0.1 s). Each metric was derived from tool 3D tip positions (x, y, z) captured by the infrared optical tracking system for all three instruments (microscissors, bipolar, aspirator). Raw coordinate streams were denoised using a fourth-order Butterworth low-pass filter after computing instantaneous velocity, acceleration, and jerk per time window. These kinematic metrics quantify motion smoothness and efficiency, while inter-instrument distance captures temporal bimanual synergy between the bipolar and aspirator. Metrics follow the computational framework described in Almansouri et al.[34] Units: mm, millimeters; s, seconds.

eTable 3. Pairwise Statistical Comparisons of Metrics and EV-ICEMS Composite Scores for All Model Architectures

Tool	Metric	AI System	*P* Value					
			Experts vs Seniors	Experts vs Juniors	Experts vs Novices	Seniors vs Juniors	Seniors vs Novices	Juniors vs Novices
Microscissors	Velocity	LSTM	0.778	0.257	0.00124	0.773	0.00919	0.0689
Microscissors	Acceleration	LSTM	0.220	0.197	0.000453	1.000	0.0686	0.0425
Microscissors	Jerk	LSTM	0.0549	0.0132	< 0.0001	0.961	0.00534	0.0118
Bipolar	Velocity	LSTM	0.0770	0.0158	0.00211	0.936	0.523	0.835
Bipolar	Acceleration	LSTM	0.0981	0.0112	< 0.0001	0.829	0.0341	0.167
Bipolar	Jerk	LSTM	0.998	0.991	0.176	0.948	0.179	0.0382
Aspirator	Velocity	LSTM	0.892	0.0394	0.0175	0.00199	0.000681	0.981
Aspirator	Acceleration	LSTM	0.826	0.000958	< 0.0001	0.00527	< 0.0001	0.00845
Aspirator	Jerk	LSTM	0.914	0.0149	< 0.0001	0.0447	< 0.0001	0.102
Bipolar-Aspirator	Inter-Instrument Distance	LSTM	0.813	0.00437	< 0.0001	0.0259	< 0.0001	0.0206
Composite	Composite	LSTM	0.499	0.0339	< 0.0001	0.452	0.00173	0.0607
Microscissors	Velocity	Transformer	1.000	0.0328	0.000683	0.0122	0.000132	0.382
Microscissors	Acceleration	Transformer	0.966	0.0273	< 0.0001	0.0498	< 0.0001	0.0102
Microscissors	Jerk	Transformer	0.999	0.0490	< 0.0001	0.0352	< 0.0001	0.0518
Bipolar	Velocity	Transformer	0.484	0.000332	< 0.0001	0.0117	0.00310	0.957
Bipolar	Acceleration	Transformer	0.358	0.00227	0.000660	0.115	0.0420	0.963
Bipolar	Jerk	Transformer	0.0384	0.000411	< 0.0001	0.380	0.0102	0.291
Aspirator	Velocity	Transformer	0.253	0.0821	0.000222	0.950	0.0303	0.0746
Aspirator	Acceleration	Transformer	0.0830	0.0168	< 0.0001	0.932	0.00186	0.00558
Aspirator	Jerk	Transformer	0.385	0.617	< 0.0001	0.956	0.000415	< 0.0001
Bipolar-Aspirator	Inter-Instrument Distance	Transformer	0.870	0.0350	0.00131	0.134	0.00534	0.509
Composite	Composite	Transformer	0.439	< 0.0001	< 0.0001	0.00399	< 0.0001	0.00133
Microscissors	Velocity	Hybrid 1	0.980	0.00388	< 0.0001	0.00489	< 0.0001	0.000132
Microscissors	Acceleration	Hybrid 1	0.449	0.00761	< 0.0001	0.197	< 0.0001	< 0.0001
Microscissors	Jerk	Hybrid 1	0.997	0.0605	< 0.0001	0.0178	< 0.0001	0.0520
Bipolar	Velocity	Hybrid 1	0.749	0.00182	0.000479	0.0157	0.00416	0.955
Bipolar	Acceleration	Hybrid 1	0.154	0.000128	< 0.0001	0.0411	0.0163	0.979
Bipolar	Jerk	Hybrid 1	0.0587	0.000141	< 0.0001	0.140	0.0123	0.692
Aspirator	Velocity	Hybrid 1	0.626	0.0249	< 0.0001	0.259	0.00140	0.122
Aspirator	Acceleration	Hybrid 1	0.888	0.0614	0.000321	0.200	0.00105	0.137
Aspirator	Jerk	Hybrid 1	0.485	0.0317	< 0.0001	0.451	0.00213	0.0724
Bipolar-Aspirator	Inter-Instrument Distance	Hybrid 1	0.880	0.0144	< 0.0001	0.0553	0.000141	0.130
Composite	Composite	Hybrid 1	0.0451	< 0.0001	< 0.0001	0.143	< 0.0001	0.000112
Microscissors	Velocity	Hybrid 2	0.959	0.314	0.00364	0.542	0.00662	0.126
Microscissors	Acceleration	Hybrid 2	0.996	0.399	0.000327	0.456	0.000180	0.00823
Microscissors	Jerk	Hybrid 2	0.381	0.117	0.000146	0.914	0.00948	0.0329

Bipolar	Velocity	Hybrid 2	0.740	0.0531	0.0293	0.318	0.201	0.991
Bipolar	Acceleration	Hybrid 2	0.0443	0.999	0.998	0.0223	0.00960	0.986
Bipolar	Jerk	Hybrid 2	0.972	0.347	0.0724	0.538	0.123	0.764
Aspirator	Velocity	Hybrid 2	0.392	0.0164	0.00136	0.394	0.0649	0.732
Aspirator	Acceleration	Hybrid 2	0.0257	0.000111	< 0.0001	0.250	< 0.0001	0.00453
Aspirator	Jerk	Hybrid 2	0.296	0.0455	0.000350	0.790	0.0344	0.195
Bipolar-Aspirator	Inter-Instrument Distance	Hybrid 2	0.937	0.0407	0.00122	0.0995	0.00269	0.448
Composite	Composite	Hybrid 2	0.0778	0.0102	< 0.0001	0.868	0.00111	0.00556

A detailed table of the *P* values associated with the heatmaps shown in **Figure 4** (main text). Values represent Tukey-adjusted *P* values following ANOVA across expertise levels (expert, senior, junior, novice). Lower values indicate stronger discrimination between groups. Values in red are significant ($P < .05$)

Aggregated across the 4-metric configuration (velocity, acceleration, jerk, and bipolar-aspirator distance)

eTable 4. Model Performance for Predictive Validity of Technical Skill Scores

Model	n	R^2	MAE	MSE	Slope	Intercept
LSTM	25	0.234	0.271	0.105	0.076	–0.327
Transformer	25	0.424	0.365	0.168	0.15	–0.577
LSTM-Transformer (Hybrid 1)	25	0.231	0.358	0.193	0.103	–0.387
Transformer-LSTM (Hybrid 2)	25	0.105	0.353	0.177	0.061	–0.26

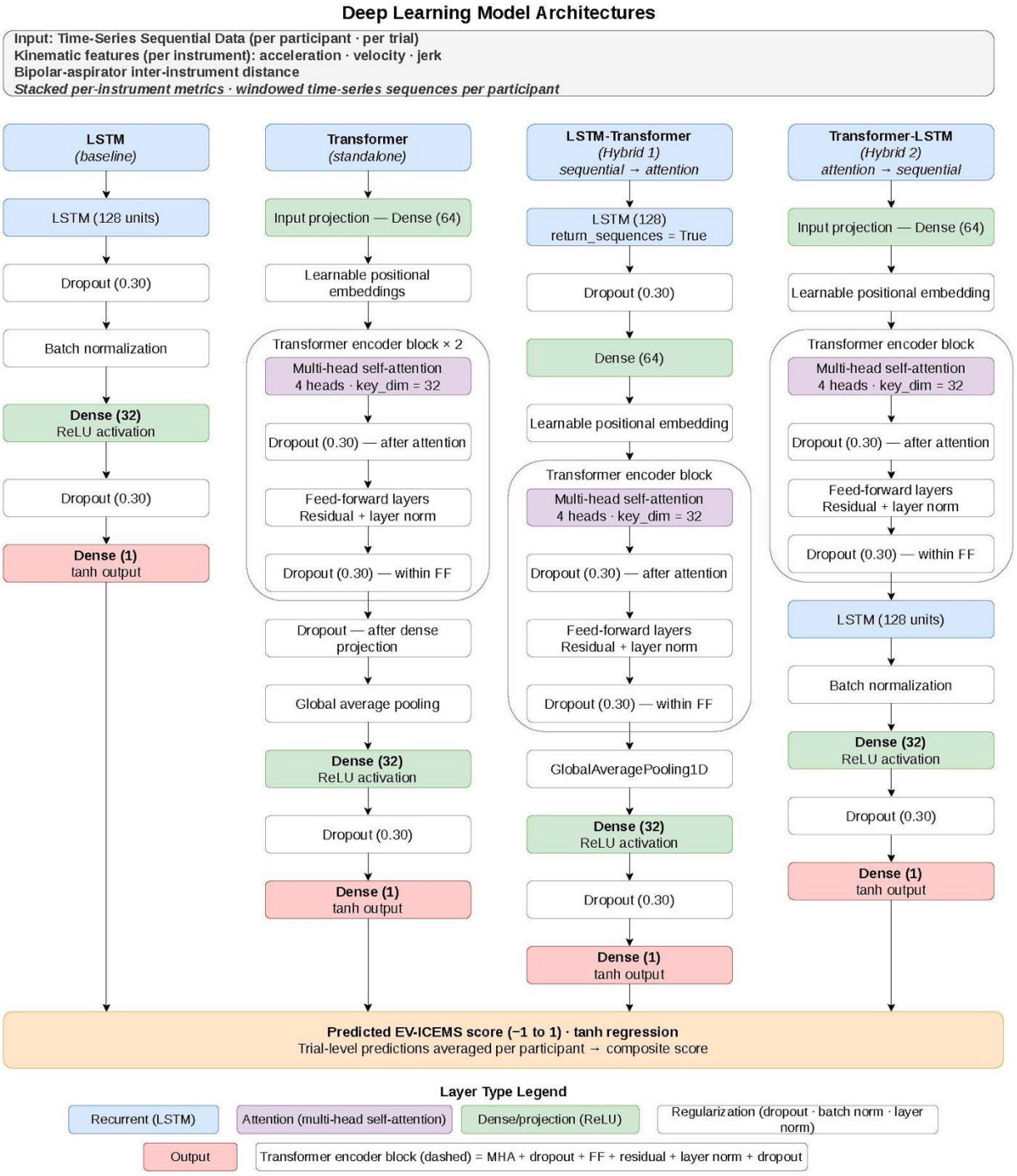
This table summarizes the predictive validity of each AI model across all junior and senior participants (n = 25). Performance is evaluated using R^2, MAE, and MSE. Higher R^2 and lower MAE/MSE indicate stronger predictive accuracy, i.e., a stronger alignment between model-derived scores and true training level. The Transformer architecture demonstrates the best overall predictive validity, achieving the highest R^2, indicating superior capture of skill progression across training years on newly tested data. It also exhibits the steepest slope, reflecting the greatest incremental improvement in predicted scores per unit increase in training level. Abbreviations: LSTM, long short-term memory; MAE, mean absolute error; MSE, mean squared error; R^2, coefficient of determination.

eTable 5. Train, Validation, and Test Performance Metrics of EV-ICEMS Models Using a 70/15/15 Split

Model	LSTM	Transformer	Hybrid 1 (LSTM-Transformer)	Hybrid 2 (Transformer-LSTM)
Train MSE	0.860	0.796	0.736	0.800
Train MAE	0.868	0.786	0.691	0.767
Train AUC	0.635	0.718	0.769	0.705
Val MSE	0.914	0.800	0.739	0.958
Val MAE	0.898	0.811	0.730	0.859
Val AUC	0.541	0.705	0.770	0.564
Test MSE	0.897	0.794	0.826	0.855
Test MAE	0.901	0.776	0.756	0.807
Test AUC	0.564	0.719	0.703	0.647
Test Accuracy	0.608	0.711	0.680	0.680

The AI architectures included various measures to prevent overfitting on the training data, such as layers of dropout regularization and the 70/15/15 training-validation-testing split, in addition to noise filtering in the pre-processing steps. To further demonstrate generalizability, this table reports training, validation, and test performance metrics across all four AI architectures (LSTM, Transformer, Hybrid 1, and Hybrid 2). Machine learning performance metrics include mean squared error (MSE), mean absolute error (MAE), and area under the receiver operating characteristic curve (AUC), as well as additional test R^2 (on novices vs. experts from the test set) and classification test accuracy. These metrics support minimal overfitting and hence, stable generalization. The Transformer demonstrated the strongest performance, with the lowest test MSE (0.794), lowest test MAE (0.776), and highest test AUC (0.719) and accuracy (0.711). These findings highlight superior discrimination and generalization on unseen data from the novice and expert groups. The hybrid models showed comparable but lower performance metrics, while the LSTM exhibited the weakest performance. These findings support our main results in the paper and further suggest that our findings highlight interpretable differences in representations of surgical expertise learned. Abbreviations: LSTM, long short-term memory; MAE, mean absolute error; MSE, mean squared error; R^2, coefficient of determination; AUC, area under the receiver operating characteristic curve.

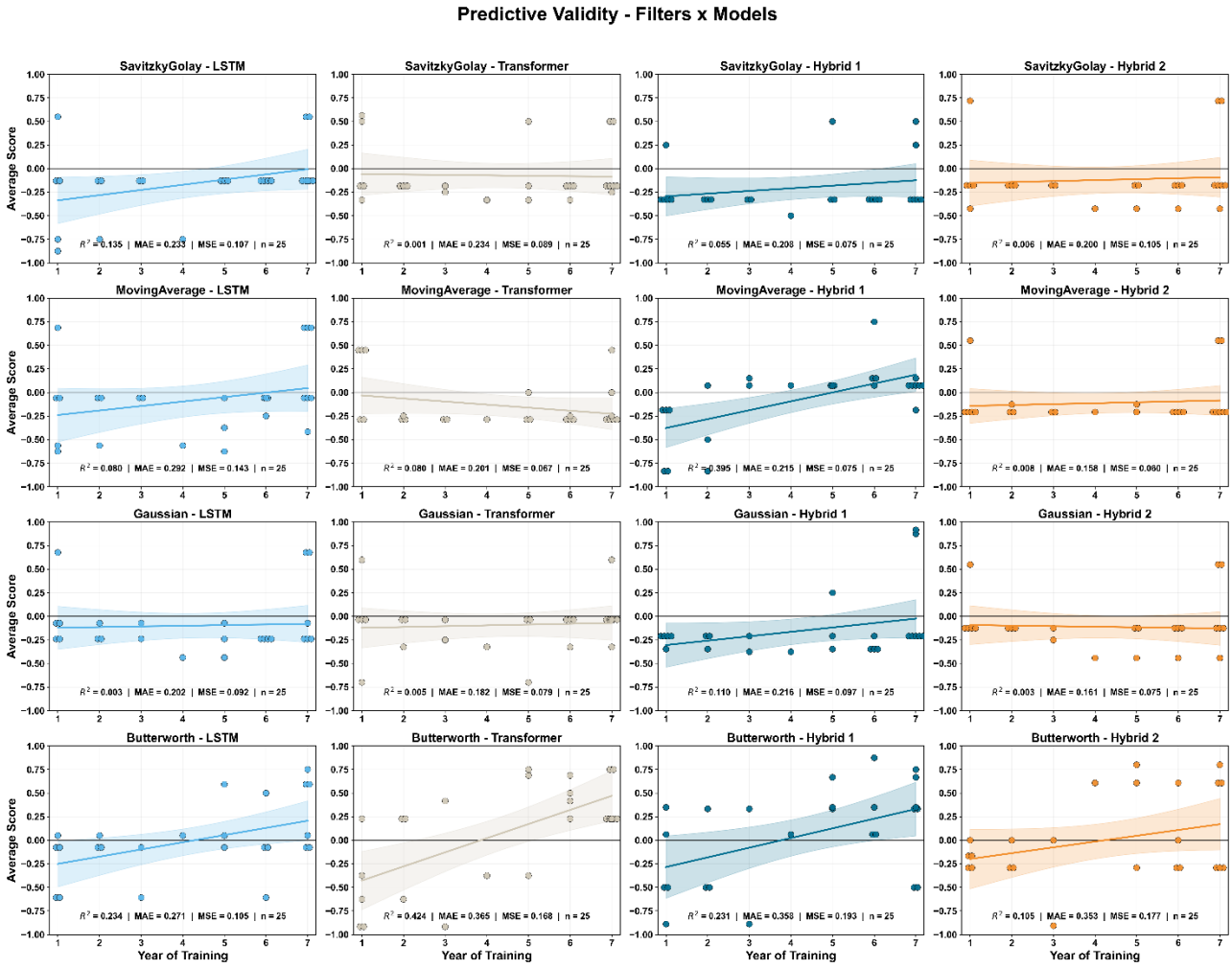
eFigure 1. Deep Learning Model Architectures for EV-ICEMS Surgical Expertise Prediction



All models process sequential time-series kinematic data sampled at 0.1 s intervals (10 Hz), including velocity, acceleration, and jerk across all instrument channels, alongside bipolar-aspirator inter-instrument distance. **(A) LSTM** (baseline): a single recurrent layer (128 units) followed by batch normalization and a dense regression head (32-ReLU, 1-tanh) with 30% dropout for regularization. **(B) Transformer**: an input projection (Dense 64) with learnable positional embeddings feeds two stacked encoder blocks, each comprising multi-head self-attention layers (4 heads, key_dim = 32), feed-forward layers with residual connections and layer normalization, and 30% dropout before the final regression scoring. **(C) LSTM-Transformer (Hybrid 1)**: sequential temporal dependencies are first captured by an LSTM unit and subsequently passed through a Transformer encoder block to compute the final

regression scores. **(D) Transformer-LSTM (Hybrid 2)**: attention-based global feature unit (Transformer encoder block) precedes an LSTM architecture, and the LSTM unit computes the final regression scores. All models apply 30% dropout within specific layers to prevent overfitting and enhance generalizability. Output scores range from –1 (novice) to 1 (expert). All models were trained identically with identical learning rates and similar hyperparameters (Adam, lr = 1×10^{-3}, MSE loss, early stopping, etc.) on expert and novice labels with 70/15/15 train/validation/test splits stratified by participant. Transformer refers to a self-attention–based sequence of model architecture. Dense denotes a fully connected neural network layer. Positional embeddings denote positional encodings added to input sequences. key_dim denotes key dimensionality in attention mechanisms. GlobalAveragePooling1D denotes global average pooling over temporal sequences. Dropout denotes stochastic regularization by randomly dropping units. Batch normalization denotes normalization of layer inputs during training. Abbreviations: LSTM, long short-term memory; ReLU, rectified linear unit; tanh, hyperbolic tangent; MHA, multi-head attention; FF, feed-forward; ICEMS, Intelligent Continuous Expertise Monitoring System.

APPENDIX B – Filters Analysis

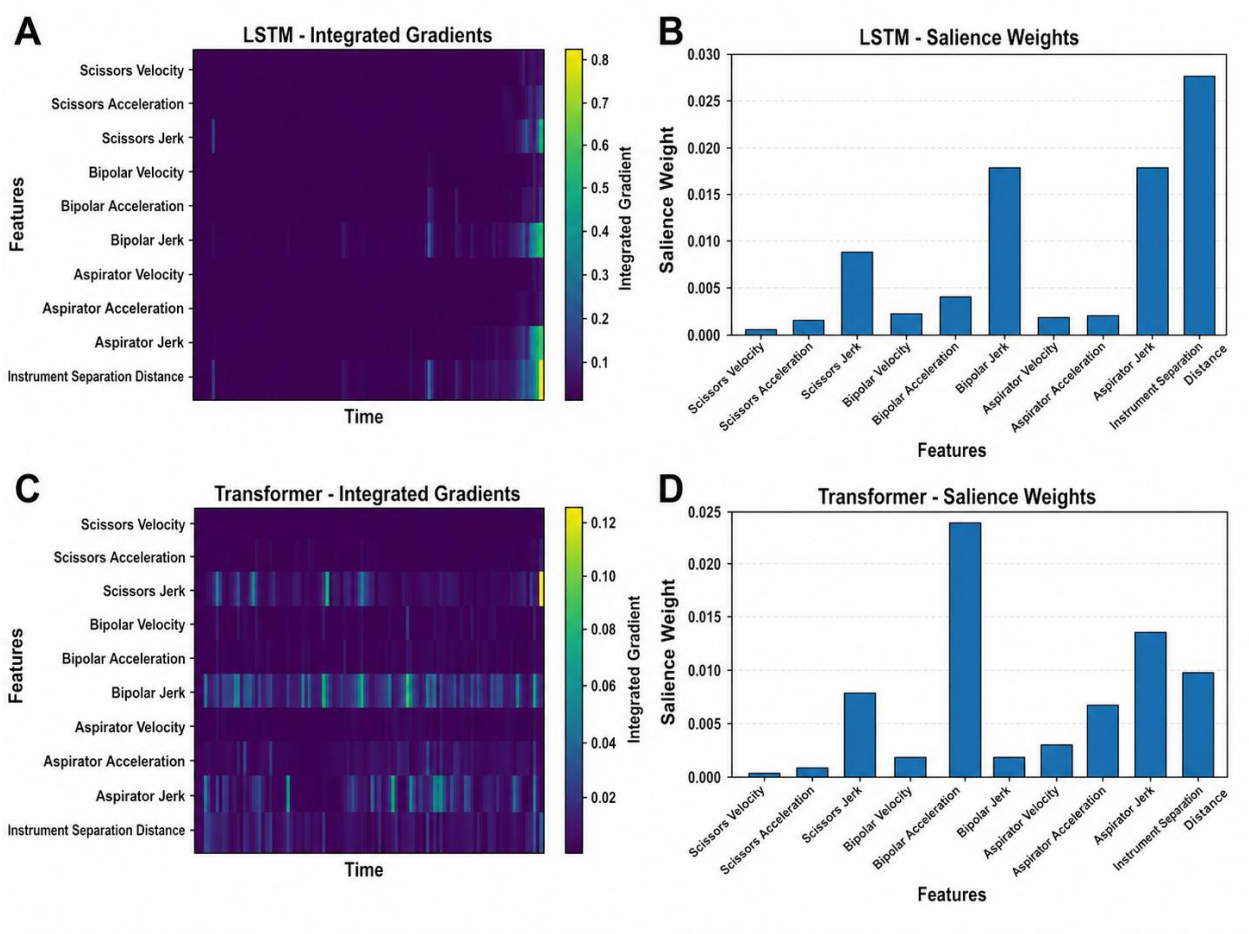


eFigure 2. Predictive Validity Across Signal Filters and AI Architectures.
Comparison of predictive validity across four distinct preprocessing filters and our AI model architectures. The Butterworth filter shows the strongest predictive validity, with the strongest predictive accuracy as denoted by the highest R^2 values across year of training, particularly for the Transformer and hybrid 1 models. In contrast, Savitzky–Golay, Moving Average, and Gaussian filtering generally show weaker or less consistent predictive trends.

Filter	Model	AUC	R2	MAE	MSE	Slope	CI_Low	CI_High
MovingAverage	Transformer	0.698	0.080	0.201	0.067	-0.032	-0.080	0.015
Gaussian	Transformer	0.679	0.005	0.182	0.079	0.009	-0.043	0.060
MovingAverage	Hybrid 2	0.676	0.008	0.158	0.060	0.010	-0.035	0.054
SavitzkyGolay	Hybrid 2	0.660	0.006	0.200	0.105	0.010	-0.049	0.070
MovingAverage	LSTM	0.657	0.080	0.292	0.143	0.048	-0.022	0.117
SavitzkyGolay	LSTM	0.637	0.135	0.233	0.107	0.055	-0.005	0.115
Gaussian	Hybrid 2	0.630	0.003	0.161	0.075	-0.006	-0.056	0.044
SavitzkyGolay	Hybrid 1	0.630	0.055	0.208	0.075	0.028	-0.022	0.079
SavitzkyGolay	Transformer	0.624	0.001	0.234	0.089	-0.005	-0.059	0.050
MovingAverage	Hybrid 1	0.590	0.395	0.215	0.075	0.094	0.044	0.145
Gaussian	LSTM	0.583	0.003	0.202	0.092	0.007	-0.049	0.063
Gaussian	Hybrid 1	0.525	0.110	0.216	0.097	0.047	-0.011	0.104

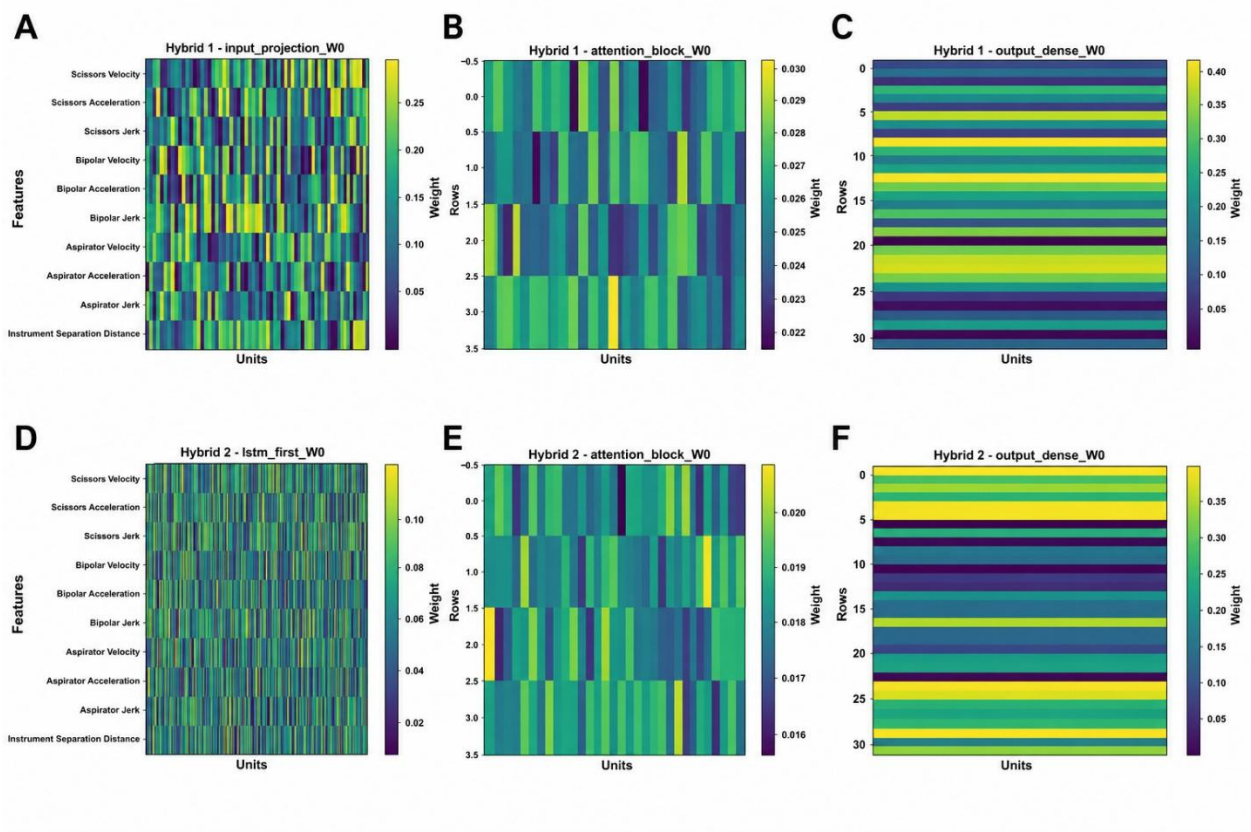
eTable 5. Comparative Predictive Performance of Non-Butterworth Filters Across AI Architectures. This table summarizes the predictive performance of the non-Butterworth preprocessing filters across four AI model architectures using discrimination and regression-based validity metrics, including AUC, R^2, MAE, MSE, slope, and slope confidence intervals. Following Butterworth, the strongest filter in terms of expert-versus-novice discrimination by AUC during the training phase is Moving Average, which contains the top non-Butterworth AUC values, led by the Transformer (AUC = 0.6981) and closely followed by Hybrid 2 (AUC = 0.6756) and LSTM (AUC = 0.6569). However, the non-Butterworth filters show weaker predictive validity, with generally lower slopes and less consistent regression across year of training compared with Butterworth. A notable exception is Hybrid 1 with Moving Average, which shows comparatively strong regression-based performance with the highest R^2 among these filters (0.3951) and a positive slope, suggesting that this filter-model combination captures training-related signal reasonably well although not as optimal.

APPENDIX C – Neural Network Architecture, Weights, and Feature Analysis



eFigure 3. Integrated gradient attribution and salience weight profiles learned by the LSTM and Transformer models from ex vivo subpial resection time-series data. A–B, LSTM temporal integrated gradient heatmap and corresponding metric-level salience weights. C–D, Transformer temporal integrated-gradient heatmap and corresponding metric-level salience weights. The heatmaps show when each kinematic tool-metric contributed most strongly across time (in AI model training), while the bar plots quantify the feature importance. The LSTM emphasized late-window instrument separation distance, aspirator jerk, and bipolar jerk, whereas the Transformer showed stronger patterns of gradient salience and learning across bipolar jerk, aspirator jerk/acceleration, and scissor jerk throughout the whole process of training.

The Transformer pattern is consistent with parallel attention-based processing. Instead of concentrating importance only at the final time points like LSTM (sequential bias), it shows more distributed, continuous activation across multiple kinematic channels, suggesting the model learned long-range temporal interdependencies and feature interactions across the full subpial resection sequence. In contrast, the LSTM appears more end-weighted, consistent with recurrent memory accumulation and stronger late-window dependence. These salience maps, integrated gradients, decoder/regression weights, and input–output attribution profiles provide an explainable-AI view of representation learning. They show which tool–metric signals the models used, not just whether the models predicted well.



eFigure 4. Architecture Layer-wise explainability maps for the Hybrid 1 and Hybrid 2 models. These layer activation maps show the learned weight structure across the input layer encoding, core sequence-processing, and output regression layers of the two hybrid models. **A–C** show Hybrid 1, including the input projection layer, an internal attention block, and the final output dense layer. **D–F** show Hybrid 2, including the first LSTM layer as the input-side sequential encoder, an internal attention block, and the final output dense layer. These network layer maps demonstrate that learning is distributed across the full network rather than being confined to a single layer, supporting end-to-end representation learning and explainable AI interpretability. The self-organized bands, gradients, and feature-specific variation suggest these are data-driven, learned patterns.