Machine learning allows expert level classification of intraoperative motor evoked potentials during neurosurgical procedures

Computers in Biology and Medicine 180 (2024) 109032

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Machine learning allows expert level classification of intraoperative motor evoked potentials during neurosurgical procedures☆,☆☆

Alessandro Boaro [a,*], Alberto Azzari [b], Federica Basaldella [c], Sonia Nunes [a], Alberto Feletti [a], Manuele Bicego [b], Francesco Sala [a]

[a] Section of Neurosurgery, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy
[b] Department of Computer Science, University of Verona, Verona, Italy
[c] Division of Neurology, University Hospital, Verona, Italy

ARTICLE INFO

Keywords:
Intraoperative neuromonitoring
Machine learning
Pattern recognition
MEP
Artificial intelligence
Neurosurgery

ABSTRACT

Objective: To develop and evaluate machine learning (ML) approaches for muscle identification using intraoperative motor evoked potentials (MEPs), and to compare their performance to human experts.
Background: There is an unseized opportunity to apply ML analytic techniques to the world of intraoperative neuromonitoring (IOM). MEPs are the ideal candidates given the importance of their correct interpretation during a surgical operation to the brain or the spine. In this work, we develop and test a set of different ML models for muscle identification using intraoperative MEPs and compare their performance to human experts. In addition, we provide a review of the available literature on current ML applications to IOM data in neurosurgery.
Methods: We trained and tested five different ML classifiers on a MEP database developed from six different muscles in patients who underwent brain or spinal cord surgery. MEPs were obtained by both transcranial (TES) and direct cortical stimulation (DCS) protocols. The models were evaluated within a single patient and on previously unseen patients, considering signals from TES and DCS both independently and mixed. Ten expert neurophysiologists classified a set of 50 randomly selected MEPs, and their performance was compared to the best performing model.
Results: A total of 25.423 MEPs were included in the study. Random Forest proved to be the best performing model with 99 % accuracy in the single patient dataset task and a 78 %–94 % accuracy range on previously unseen patients. The model performance was maximized by representing MEPs as a set of features typically employed in signal processing compared to traditional neurophysiological parameters. The classification ability of the Random Forest model between six different muscles and across different MEP acquisition modalities (79 %) significantly exceeded that of human experts (mean 48 %).
Conclusions: Carefully selected ML models proved to have reliable capacity of extracting meaningful information to classify intraoperative MEPs using a limited number of features, proving robustness across patients and signal acquisition modalities, outperforming human experts, and with the potential to act as decision support systems to the IOM team. Such encouraging results lay the path to further explore the underlying nature of clinically important signals, with the aim to continue to produce useful applications to make surgeries safer and more efficient.

1. Introduction

The monitoring of motor evoked potentials (MEPs) during a surgical operation on the brain or the spine is a well-established technique for real-time functional integrity assessment [1,2,3]. Intraoperative monitoring (IOM) is performed by connecting a standardized set of muscles of the anesthetized patient to a monitoring machine, which enables the

☆ Previous Presentations: Portions of this work have been presented in form of oral presentation or poster at the 2022 EANS Annual Congress in Belgrade (Serbia), the 2022 ISIN Annual Congress in Chicago (USA), the 2022 SINch Annual Congress in Naples (Italy), and at the EANS 'Bridging Neurosciences and Neurosurgery: New Frontiers in Intraoperative Neurophysiology' Symposium in Verona (Italy), 2023.☆☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author.
E-mail address: alessandro.boaro@univr.it (A. Boaro).

https://doi.org/10.1016/j.compbiomed.2024.109032
Received 12 January 2024; Received in revised form 7 August 2024; Accepted 13 August 2024
Available online 21 August 2024

Abbreviations

ML	Machine learning
TES	Transcranial electrical stimulation
DCS	Direct cortical stimulation
IOM	Intraoperative neuromonitoring
TSFRESH	Time Series FeatuRe Extraction on basis of Scalable Hypothesis
NN	Nearest Neighbor
linSVM	linear support vector machine
rbfSVM	radial basis function support vector machine
RF	Random Forest

electrical stimulation of the neural pathways controlling the muscles and the visualization of the responses on a dedicated screen (Fig. 1). Classical stimulation methods include transcranial electrical stimulation (TES), in which the stimulation is delivered through scalp electrodes, and direct cortical stimulation (DCS), in which the stimulation is delivered directly on the exposed cerebral cortex using strip electrodes [4–6]. In both cases, the response is recorded at the level of target muscles using subdermal electrodes. The correct implementation of the monitoring settings and interpretation of the motor outputs, with regards to ongoing surgical events, are essential to accurately warn of signal changes that could lead to clinically meaningful deficits such as the inability to walk, the inability to talk or the loss of hand motility [2,3,7–9] Current interpretation of MEPs relies solely on the semi-quantitative evaluation of a handful of parameters by experienced neurophysiologists. These professionals mostly evaluate MEP amplitude and the magnitude of its reduction, along with additional elements as increase in latency, threshold elevation and morphology simplification that can be considered and are variably employed [8,10–13].

Nonetheless, MEPs are complex biological signals produced by summation of multiple electrical components, that present a certain variability within and between subjects and even from one stimuli to the next one, due to endogenous brain states or exogenous interventions [14–16]. Such complexity and variability can hardly be captured and effectively conveyed by a handful of parameters, making the real-time monitoring of motor functions during a neurosurgical operation quite challenging (Fig. 2). In addition, despite the complexity and the number of connections happening between the patient and the monitoring machine (Fig. 3), currently there is no safety checklist to confirm the relationship between each muscle and its label, which is manually assigned on the machine. The muscle identification task, upon which depends the reliability of the signal interpretation during the whole surgical procedure, is currently entirely entrusted to the expertise of the neurophysiologist. Such a task becomes of even higher difficulty in consideration of the growing occurrence of surgical environments in which a single neurophysiologist is following multiple patients in different operative rooms at the same time.

Recent development in data analysis techniques, specifically in the form of Machine Learning (ML) approaches applied to medical data such as imaging, histology data, biological signals and biometric data, proved the ability to efficiently extract and summarize clinically meaningful information [17–21]. With specific regards to the analysis of electrical biological signals, some successful and clinically significant examples include the development of classification algorithms to recognize different types of cardiac arrhythmias from EKG, or the ability to recognize seizures from EEG or again using ML models to detect muscular activity from surface EMG used in the programming of robotic prosthesics [22–24]

Regardless, there is a currently unseized opportunity in applying these powerful analytic techniques to the world of intraoperative neuromonitoring (IOM), with an unmet potential to deepen the knowledge about the biological signals involved, to increase the safety of surgical and medical procedures, and to refine diagnostic and warning criteria.

In this work, we developed and tested a set of different ML models in the identification of different muscles by analyzing intraoperative MEPs and compared their classification performance to that of human experts. Finally, we conducted a review of the available literature on current ML applications on IOM data in neurosurgery, in order to provide adequate context to our work and assess the overall clinical readiness of these technologies.

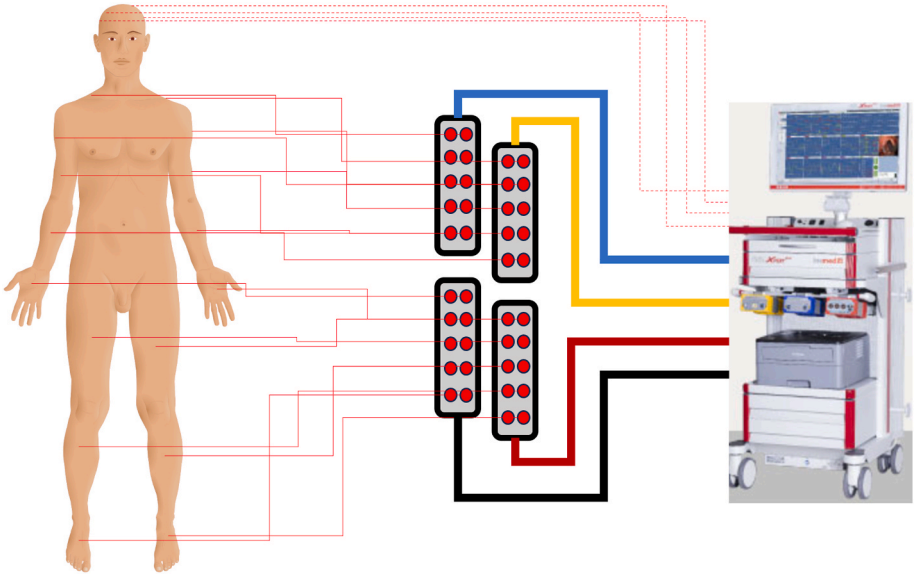

Fig. 1. Schematic drawing of a standard of care IOM setting. The patient (left) is connected to the monitoring machine (right) by means of a large number of cables (continuous and dashed red thin lines), interposing amplification boxes (center) and derivation cables (four colored thick lines). The stimulation cables (dashed thin red lines) provide the stimulation current to the brain and the MEPs are registered at the muscle level and sent back to the monitoring machine through the acquisition cables (continuous thin red lines).

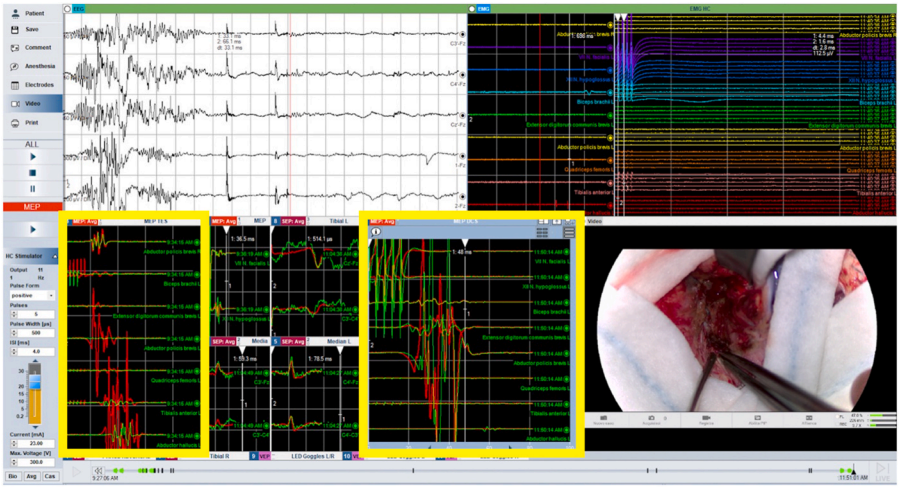

Fig. 2. Visualization of signal traces on monitor. The picture depicts the complexity of the signal traces as they are presented to the eye of the neurophysiologist during a surgical operation. The yellow squares identify the two boxes where the MEPs are visualized.

2. Methods

2.1. Patients' selection

Patients were retrospectively selected from the database of the neurosurgery department at the Verona Regional Hospital based on the following criteria: a) clinical and radiological diagnosis of brain or spinal cord lesions with indication for surgical treatment, b) employment of TES, DCS or both as intraoperative MEP monitoring techniques, c) inclusion of biceps brachii (bb), extensor digitorum communis brevis (edcb), abductor pollicis brevis (apb), quadriceps femoris (qf), tibialis anterior (ta) and abductor hallucis (ah) among the muscles monitored d) presence of monitorable responses. Pediatric patients were excluded. All procedures were performed in compliance with institutional guidelines and have been approved by the appropriate institutional committee. The study obtained IRB approval with protocol ID: CRMS 23038 on the June 05, 2024.

2.2. MEP acquisition

For the TES protocol, MEPs were evoked by application of transcranial anodal electrical stimulation with corkscrew electrodes placed on the scalp at C1 and C2 by 10–20 International System. For the DCS protocol, MEPs were evoked by application of anodal electrical stimulation directly on the cortex through strip electrode as anode and a corkscrew electrode as cathode, placed on the scalp in Fz position by 10–20 International System.

High frequency short train technique was used, by application of a train of five pulses with an interstimulus interval of 4 ms for TES and 2 ms for DCS. Intensity ranged between 50 and 150 mA for TES and 5–20 mA for DCS.

MEPs were recorded from the same target muscles in all patients, with pairs of subdermal needle electrodes placed into the muscle belly of biceps brachii, extensor digitorum communis brevis, abductor pollicis brevis for upper limbs and quadriceps femoris, tibialis anterior and abductor hallucis for lower limbs.

2.3. Database creation

The MEP database was created by extracting and selecting all the available MEPs from each single patient by an experienced neurophysiologist (F.B.) and an experienced neurosurgeon (A.B.). Signals with high grade of noise or silent responses were excluded as well as signals with latency not corresponding to each specific muscle physiological range. The database was then duplicated in order to have one database where each MEP was kept in its entirety (whole signal database) and another database where the latency (the time occurring from stimulation to MEP appearance) was removed and only the MEP waveform was kept (no latency database) (schematic examples of signals from each database in Fig. 4). For both databases, we produced five different signal representations and specifically: RAW (unprocessed signal), NORM (signals are normalized between -1 and $+1$), TRAD (each signal is represented as a set of 8 traditional neurophysiological features: amplitude, area, duration, thickness, size index, number of phases, number of turns, number of satellites) [25], TSFRESH (each signal is represented as a set of 789 features produced by a time series feature extraction model), TSFRESH_FS (each signal is represented as a selection of the 10 most meaningful features from TSFRESH) [26,27]. To get this last representation, we first employed the feature selection method proposed in TSFRESH, which returned a set of 182 features; subsequently, in order to have a number of features comparable with the set of traditional features used by neurophysiologists (TRAD), we refined the selection by employing the Recursive Feature Elimination approach [28,29]. The pipline just described to produce the TSFRESH_FS set, was employed twice and separately for the whole signal database and the no latency database, in order to properly take into account the presence/absence of the signal latency in the feature extraction process.

2.4. Machine learning models training and testing

A total of 5 different types of Machine Learning classifiers were trained and tested: k-Nearest Neighbor with $k = 1$ (NN), k-Nearest Neighbor with $k = 10$ (kNN), linear support vector machine (linSVM), radial basis function support vector machine (rbfSVM), Random Forest (RF) [30–32]. The k-Nearest Neighbors (k-NN) algorithm is an instance-based learning method [31] that classifies a data point based on the classes of its neighbors. It assumes that similar instances are in close proximity within the feature space. For a given data point, k-NN identifies the 'k' closest training examples and assigns the most frequent label among these neighbors to the data point. In our experiments, we employed the standard Euclidean Distance, using two values for k: $k = 1$ (i.e. the Nearest Neighbor rule) and $k = 10$. Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks [33]. SVMs face classification by identifying the hyperplane that best separates data points of different classes; this is obtained by maximizing the so-called margin. To get a non-linear decision function the kernel-trick is used, which permits to project the data into a higher-dimensional space where classes can be more easily separated. In

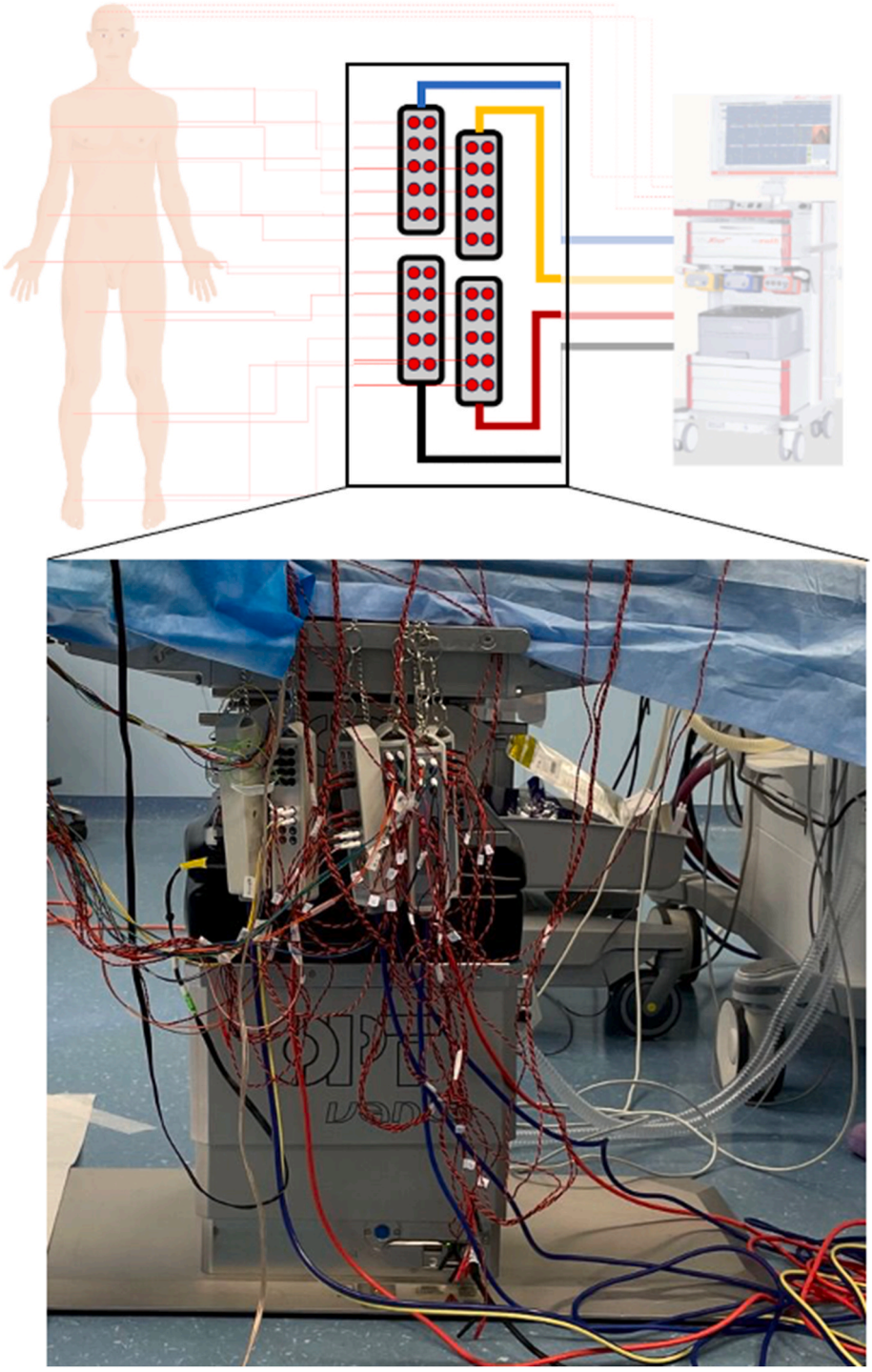

Fig. 3. Photograph of amplification boxes and their connections. This photograph clearly depicts the complexity of the cable connections that must be put in place to perform a standard intraoperative monitoring of motor functions.

our study we employed two versions of SVM, each one corresponding to a different kernel: linSVM, i.e. a SVM with a linear kernel, and rbf-SVM, i.e. a SVM in which a Radial Basis Function Kernel is employed. In both cases, we set the cost optimization parameter C (representing a measure of the compromise between generalization and errors in the training set) via 5-fold cross validation on the training set. In a similar way we also set the parameter sigma for the rbf kernel. Random Forests (RF) represent an approach for classification and regression [32] based on an ensemble learning strategy. Specifically, a RF constructs multiple decision trees and merges their outputs for improved accuracy and stability. To ensure diversity among trees, each tree in the forest is trained on a different bootstrapped sample of the data, by using a different random subset of features or by employing other randomization mechanisms. For classifying an object, the algorithm exploits a majority voting scheme among the results of the different trees. In our experiments we employed 100 trees, each one trained using the classic GINI criterion.

A total of three classification tasks of increasing complexity were defined.

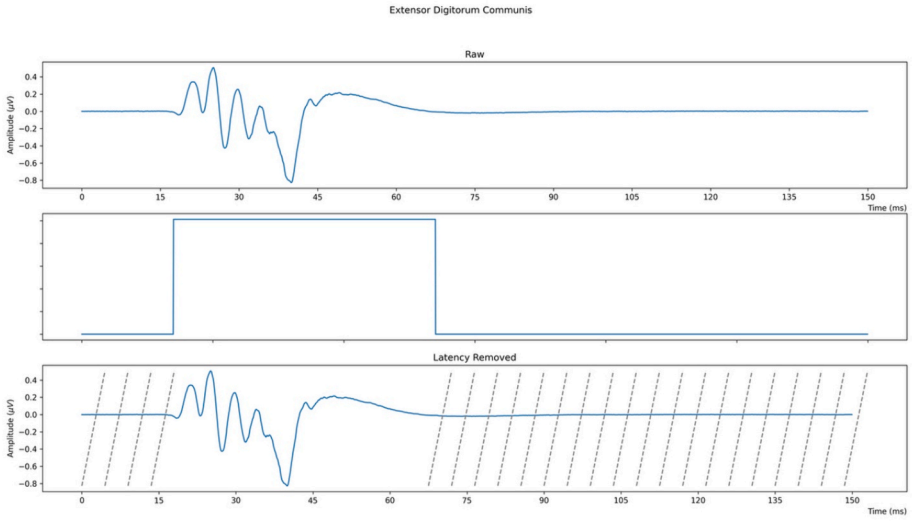

Fig. 4. Whole signal and no latency MEP example. On top, the first plot shows an example of a raw signal as considered in the whole signal database. The second plot shows the interval from which the signal waveform is extracted, and the third plot shows the final extraction result as an example of a signal as considered in the no latency database. The X axis of each plot represents time in ms (milliseconds), the Y axis of each plot represents MEP amplitude in μV (microvolts).

- **Task 1.** Intra-patient. In this task the models were trained on the signals of a single patient and were tested on their ability to correctly classify the MEPs of the same patient. MEPs from TES and DCS protocols were considered separately.
- **Task 2.** Inter-patient. In this task the models were trained on the signals of all patients except one and were tested on their ability to correctly classify the MEPs of the patient not considered. As not all the muscle categories were represented for all patients, two versions of task 2 have been deployed, one in which the number of patients was maximized and two muscles of the upper limb included, and another in which the number of muscles (located in both the upper and lower limbs) was maximized. MEPs from TES and DCS protocols were considered separately.
- **Task 3.** Inter-protocol. In this task signals from TES and DCS protocols were mixed, and the models were tested in their classification ability irrespective of the monitoring protocol used.

Each model was trained and tested on all five signal representations independently.

Each task was conducted on both the whole signal and the no latency databases.

The performance of each classifier was evaluated using cross-validation techniques and specifically, repeated stratified k-fold cross-validation (in task 2.1 k is equal to 25 in DCS and 10 in TES, in task 2.2 k is equal to 15 in DCS and 9 in TES) and stratified k-group cross-validation (for both task 1 and task 3 we set k to 5 and r to 10), in order to provide a more reliable estimate of each model's performance on unseen data as well as to maximize model robustness and generalizability [34]. The performances have been measured and presented in form of balanced accuracy (reported in the text and in Tables 4–6), F1 score and Matthews's correlation coefficient (MCC) (reported in Tables 4–6) [35–37].

2.5 model vs expert comparison

Ten expert neurophysiologists were asked to independently classify a set of 50 MEPs randomly selected from the whole signal database. Their classification performance expressed as the percentage of correctly classified signals, was compared to the one from the best performing model on the most difficult and generalizable task (task 3).

Table 1
Patients' characteristics.

Total	54
Age (mean, years)	58
Gender (M:F)	1:1
Pathology	
Brain-tumor	24
Brain-vascular	8
Spine-tumor	17
Spine-fracture	1
Spine degenerative	2
Spine-malformation	2
IOM protocol	
TES	26
DCS	28

Table 2
MEP database. The table reports the databases composition specific to muscle and IOM protocol.

	TES total	DCS total	Task 2.1 TES	Task 2.1 DCS	Task 2.2 TES	Task 2.2 DCS
Biceps Brachii	2438	2208	/	/	2167	904
Extensor Digitorum Communis	2714	3128	2714	3101	1833	1601
Abductor Pollicis Brevis	3433	3781	2971	3664	2179	1865
Quadriceps Femoris	1593	312	/	/	/	/
Tibialis Anterior	2311	702	/	/	1189	/
Abductor Hallucis	1987	816	/	/	/	804
Total	**14476**	**10947**	**5685**	**6765**	**7368**	**5174**

2.6 Literature review

We systematically reviewed the available literature to assess current applications of artificial intelligence methods in the interpretation of IOM data in neurosurgery. The literature review was conducted in accordance with the 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement [38]. MEDLINE

Table 3
TSFRESH_FS features. The table reports the TSFRESH_FS composition for the whole signal database and the no latency databases.

Position	Whole signal	No latency
1	ratio beyond r sigma (r = 1)	autocorrelation (lag = 8)
2	autocorrelation (lag = 9)	agg. autocorrelation(aggtype = median, maxlag = 40)
3	index mass quantile (q = 0.1)	agg. autocorrelation (aggtype = mean, maxlag = 40)
4	spkt welch density (coeff = 8)	change quantiles (qh = 0.2, ql = 0.0)
5	lempel-ziv complexity (bins = 100)	ar coefficient (coeff = 3, k = 10)
6	index mass quantile (q = 0.2)	agg. linear trend (chunklen = 10)
7	index mass quantile (q = 0.3)	fft coefficient (attr = image, coeff = 1)
8	index mass quantile (q = 0.4)	index mass quantile (q = 0.1)
9	energy ratio by chunks (n.seg = 10, seg.foc = 0)	energy ratio by chunks (n.seg = 10, seg.foc = 0)
10	fft aggregated (aggtype = variance)	cwt coefficients (coeff = 10)

r: ratio, q: quantile, agg: aggregator, aggtype: aggregator type, spkt: spectral, coeff: coefficient, qh: higher quantile, ql: lower quantile, ar: autoregressive, k: maximum lag, chunklen: length of chunk, attr: attribute, n.seg: number of segments, seg.foc: segment focus, cwt: continuous wavelet transform.

Table 4
Task 1 performances. The table reports the peak performances per model and related signal representation in Task 1. The highest performances are reported in bold. The standard error of the mean for accuracy, f1 score and mcc are respectively all lower than 0.004, 0.004, 0.005.

	TES		DCS	
	Whole signal	No latency	Whole signal	No latency
Accuracy				
kNN (1)	96 % (raw/ts_fs)	93 % (raw)	97 % (raw/norm)	97 % (raw/norm)
kNN (10)	91 % (ts_fs)	82 % (norm)	90 % (ts_fs)	88 % (norm)
linSVM	93 % (ts)	88 % (raw/norm)	97 % (norm/ts)	96 % (norm)
rbfSVM	89 % (norm)	89 % (norm)	93 % (raw/norm)	95 % (norm)
RF	**99 % (ts)**	**97 % (ts)**	**99 % (ts)**	**99 % (ts)**
F1 score				
kNN (1)	97 % (raw/norm/ts_fs)	95 % (raw/norm)	98 % (raw/norm)	98 % (norm)
kNN (10)	95 % (ts_fs)	89 % (norm)	94 % (ts_fs)	95 % (norm)
linSVM	96 % (ts)	91 % (raw/norm)	98 % (norm/ts)	98 % (norm)
rbfSVM	95 % (norm)	95 % (norm)	97 % (raw/norm)	98 % (norm)
RF	**99 % (ts)**	**98 % (ts)**	**99 % (raw/ts/ts_fs)**	**99 % (raw/ts/ts_fs)**
MCC				
kNN (1)	96 % (raw/ts_fs)	92 % (raw/norm)	97 % (raw/norm)	97 % (norm)
kNN (10)	93 % (ts_fs)	85 % (norm)	92 % (ts_fs)	92 % (norm)
linSVM	94 % (ts)	87 % (norm)	97 % (norm/ts)	97 % (norm)
rbfSVM	92 % (norm/ts_fs)	92 % (norm)	96 % (raw/norm)	97 % (norm)
RF	**99 % (ts)**	**97 % (ts)**	**99 % (ts)**	**99 % (ts)**

ts: TSFRESH, ts_fs: TSFRESH FS.

(PubMed), Embase, and Cochrane databases were searched using keywords related to intraoperative neuromonitoring techniques (MEP, SSEP, VEP, EMG) in combination to artificial intelligence-related terms (machine learning, deep learning, artificial intelligence). Relevant studies identified in the bibliographies of the reviewed papers were also included. Duplicate publications, non-English language papers, reviews,

Table 5
Task 2 performances. The table reports the peak performances per model and related signal representation in Task 2. The highest performances are reported in bold. In Task 2.1, the standard error of the mean for accuracy, f1 score and mcc are respectively all lower than 0.003, 0.003, 0.005. In Task 2.2, the standard error of the mean for accuracy, f1 score and mcc are respectively all lower than 0.003, 0.004, 0.004.

TASK 2.1	TES		DCS	
	Whole signal	No latency	Whole signal	No latency
Accuracy				
kNN (1)	85 % (ts_fs)	76 % (raw)	89 % (ts_fs)	73 % (raw)
kNN (10)	87 % (ts_fs)	76 % (norm)	90 % (ts_fs)	73 % (norm)
linSVM	90 % (ts_fs)	89 % (ts_fs)	**92 % (ts_fs)**	86 % (ts_fs)
rbfSVM	90 % (norm)	80 % (norm)	90 % (ts_fs)	76 % (raw)
RF	**92 % (ts_fs)**	**94 % (ts)**	90 % (ts_fs)	**90 % (ts)**
F1 score				
kNN (1)	82 % (ts_fs)	71 % (raw)	90 % (ts_fs)	73 % (raw)
kNN (10)	85 % (ts_fs)	71 % (norm)	91 % (ts_fs)	74 % (raw)
linSVM	88 % (ts_fs)	87 % (ts_fs)	**93 % (ts_fs)**	85 % (ts_fs)
rbfSVM	88 % (ts_fs)	73 % (norm)	91 % (ts_fs)	78 % (raw)
RF	**89 % (raw/ts_fs)**	**91 % (ts)**	92 % (ts_fs)	**91 % (ts_fs)**
MCC				
kNN (1)	68 % (ts_fs)	48 % (raw)	78 % (ts_fs)	44 % (raw)
kNN (10)	76 % (ts_fs)	49 % (norm)	80 % (ts_fs)	47 % (raw)
linSVM	82 % (ts_fs)	79 % (ts_fs)	**85 % (ts_fs)**	73 % (ts_fs)
rbfSVM	82 % (ts_fs)	59 % (norm)	81 % (ts_fs)	56 % (raw)
RF	**86 % (/ts_fs)**	**89 % (ts)**	83 % (ts_fs)	**82 % (ts/ts_fs)**

TASK 2.2	TES		DCS	
	Whole signal	No latency	Whole signal	No latency
Accuracy				
kNN (1)	76 % (ts_fs)	56 % (raw)	75 % (ts_fs)	49 % (norm)
kNN (10)	78 % (ts_fs)	55 % (norm)	73 % (ts_fs)	47 % (norm)
linSVM	76 % (ts_fs)	70 % (ts_fs)	70 % (ts_fs)	54 % (ts_fs)
rbfSVM	79 % (ts_fs)	60 % (norm)	76 % (ts_fs)	50 % (norm)
RF	**84 % (ts)**	**78 % (ts)**	**79 % (ts_fs)**	**61 % (ts)**
F1 score				
kNN (1)	72 % (ts_fs)	58 % (norm)	70 % (ts_fs)	49 % (raw)
kNN (10)	74 % (ts_fs)	60 % (norm)	67 % (ts_fs)	49 % (ts)
linSVM	71 % (ts_fs)	72 % (ts_fs)	62 % (ts_fs)	54 % (ts_fs)
rbfSVM	74 % (ts_fs)	63 % (norm)	71 % (ts_fs)	47 % (norm)
RF	**85 % (ts)**	**81 % (ts)**	**73 % (ts_fs)**	**64 % (ts)**
MCC				
kNN (1)	65 % (ts_fs)	41 % (raw)	62 % (ts_fs)	30 % (norm)
kNN (10)	68 % (ts_fs)	44 % (norm)	60 % (ts_fs)	29 % (norm)
linSVM	67 % (ts_fs)	62 % (ts_fs)	58 % (ts_fs)	43 % (ts_fs)
rbfSVM	70 % (ts_fs)	48 % (norm)	65 % (ts_fs)	32%norm)
RF	**80 % (ts)**	**74 % (ts)**	**68 % (ts_fs)**	**56 % (ts)**

ts: TSFRESH, ts_fs: TSFRESH FS.

and case reports were excluded. Two independent authors (A.B., S.N.) screened the titles and abstracts of articles against the inclusion and exclusion criteria. Subsequently, full texts were reviewed against eligibility criteria for final selection. Any disagreements between the authors were resolved by discussion.

Extracted data included (1) study information; (2) type of IOM technique employed; (3) patient population characteristics; (4) type of surgery; (5) type of ML approach employed; (6) ML model aim; (7) variables included in the model; (8) model performance; (9) interpretability assessment; (10) presence of prospective validation; (11) comparison to human/traditional methods performance. The findings are presented in the form of a narrative review in the context of the discussion of our results.

3 Results

3.1 Patients and MEPs characteristics

We included a total of 54 patients who underwent either oncological, vascular or trauma surgery to the brain or the spine (Table 1). TES

Table 6
Task 3 performances. The table reports the peak performances per model and related signal representation in Task 3. The highest performances are reported in bold. The standard error of the mean for accuracy, f1 score and mcc are respectively all lower than 0.001, 0.001, 0.001.

	Whole signal	No latency
Accuracy		
kNN (1)	70 % (ts_fs)	44 % (raw)
kNN (10)	72 % (ts_fs)	43 % (norm)
linSVM	64 % (ts_fs)	53 % (ts_fs)
rfbSVM	72 % (ts_fs)	49 % (norm)
RF	**79 % (ts)**	**66 % (ts)**
F1 score		
kNN (1)	71 % (ts_fs)	45 % (raw)
kNN (10)	75 % (ts_fs)	46 % (norm)
linSVM	67 % (ts_fs)	57 % (ts_fs)
rfbSVM	76 % (ts_fs)	51 % (norm)
RF	**81 % (ts)**	**68 % (ts)**
MCC		
kNN (1)	65 % (ts_fs)	31 % (raw)
kNN (10)	70 % (ts_fs)	32 % (raw/norm)
linSVM	63 % (ts_fs)	49 % (ts_fs)
rfbSVM	71 % (ts_fs)	39 % (norm)
RF	**77 % (ts)**	**61 % (ts)**

ts: TSFRESH, ts_fs: TSFRESH FS.

protocol was performed in 26 patients and DCS protocol was employed in 28 patients. We collected a total of 25423 MEPs (14476 MEPs of TES protocol and 10947 MEPs of DCS protocol) variably distributed between different patients and muscles (Table 2). Each feature included in TSFRESH representation was part of one of the following categories: signal summary statistics (e.g., max, min, length, etc.) characteristics of sample distribution (e.g., symmetry, entropy, energy, etc.), regression and correlation models (e.g., autoregression coefficients, etc.), observed dynamics in the time domain (e.g., index of the first max, first derivative, maximal length over the mean, etc.) or observed dynamics in the frequency domain (e.g., frequency analysis, Fourier coefficients, wavelets coefficients, etc.). The top 10 features composing the TSFRESH_FS representation are reported in Table 3.

3.2 Models' performances

3.2.1. Intra-patient classification – task 1

In the intra-patient classification task, with regards to the whole signal database, the classification performances expressed as balanced accuracy reached a maximum of 99 % both in the TES and in the DCS groups using the RF model on the TSFRESH representation (Table 4). The lowest performances were measured for the rbfSVM on the TRAD representation for the TES group (49 %) and again on the TSFRESH representation for the DCS group (58 %). With regards to the no latency database, the RF model applied to the TSFRESH representation proved to be the best with 99 % accuracy in the DCS group and 97 % in the TES group (Table 4). The lowest performing model was the rbfSVM applied to the TSFRESH_FS representation with 48 % in the TES group and 55 % in the DCS group.

3.2.2. Inter-patient classification – task 2.1 (number of patients is maximized)

In the first inter-patient classification task we considered a total of 41 patients and two muscle classes (extensor digitorum communis brevis and abductor pollicis brevis). In the TES group the RF produced the highest performance both in the whole signal and no latency databases respectively with 92 % for the TSFRESH_FS and 94 % for the TSFRESH representations. In the DCS group all the models performed similarly high in the whole signal database with peak performances of 89–92 % on the TSFRESH_FS representation. In the no latency database, the highest performance was obtained by the RF on TSFRESH and TSFRESH_FS with 90 % accuracy. (Table 5). The lowest performances were recorded for the linSVM in the RAW and NORM representations in all the instances (43 %–47 %) (Supplementary file 1).

Confusion matrices are provided in Fig. 5 for the best perfoming models and related signal representations, where comparable results in terms of classification accuracy are evident across muscles, iom modalities and presence/absence of signal latency.

3.2.3. Inter-patient classification – task 2.2 (number of muscles is maximized)

In the second inter-patient classification task we considered a total of 21 patients and performed a four-muscle classification including biceps brachii, extensor digitorum communis brevis, abductor pollicis brevis and abductor hallucis. In the TES group, the highest performances were reached by RF both in the whole signal and no latency databases, respectively in the TSFRESH representation (84 %) and in the TSFRESH/TSFRESH_FS representations (78 %). Similarly, in the DCS group, the highest performances were reached by RF both in the whole signal and no latency databases, respectively in the TSFRESH_FS representation (79 %) and in the TSFRESH representation (61 %) (Table 5). The lowest performances were obtained by linSVM in both groups and databases (DCS/whole signal/RAW 27 %, DCS/no latency/RAW 21 %, TES/whole signal/TSFRESH 22 %, TES/no latency/TSFRESH 22 %) (Supplementary file 1). Confusion matrices are provided in Fig. 6 for the best performing models and related signal representations.

3.2.4. Inter-protocol classification – task 3

In the final task, signals from TES and DCS groups were combined, and all six muscles included. In the whole signal database, the best performing model was RF on TSFRESH, RAW and TSFRESH_FS representations with 77–79 % accuracies (Table 6), while the lowest was the linSVM on the TSFRESH representation with only 16 % accuracy. In the no latency database, the best performing model remained the RF with 65–66 % on the TSFRESH and TSFRESH_FS representations (Table 6), while the lowest performing model was the linSVM with 18 % on the RAW representation (Supplementary file 1).

In all three tasks, variations between balanced accuracy, F1 weighted, and MCC metrics are negligible, indicating consistent model performance across measures.

3.3 Expert classification performance

The muscle identification task performed by the ten experts over 50 randomly selected MEPs reached an overall accuracy of 47.4 ± 11.9 % on average, while the best performing model on the most difficult task, which included signals from mixed protocols and produced by all six muscle classes, was Random Forest with an accuracy of 79 % (Fig. 7).

3.4 Current applications of ML to IOM data in neurosurgery

A total of 3356 studies were identified during the search. Of those, 3301 were excluded in title and abstract screening. After full text review, 48 studies were excluded because either no surgical application was employed, or no intraoperative data were used, or the target population of the study was not human. A total of 7 studies were included in the final data extraction phase (Fig. 8, Table 7). Two studies focused on EMG data [39,40], one on VEP data [41], one on ECoG and DBS data [42], one on SSEP data [43] and two on MEPs [44,45]. Five papers employed some form of supervised learning approach while two out of six applied unsupervised clustering methods. Four papers developed models aimed at providing an estimate or prediction of clinically relevant post-operative outcome, one aimed at refining IOM signals by removing stimulation artifacts, one focused on transcranial MEP classification, and one aimed at developing a tool for the automated intra-operative central sulcus delineation. Four papers provided model interpretability assessments, mainly in form of visual maps and six models out of seven compared the models' performance to that of human or traditional methods. No model

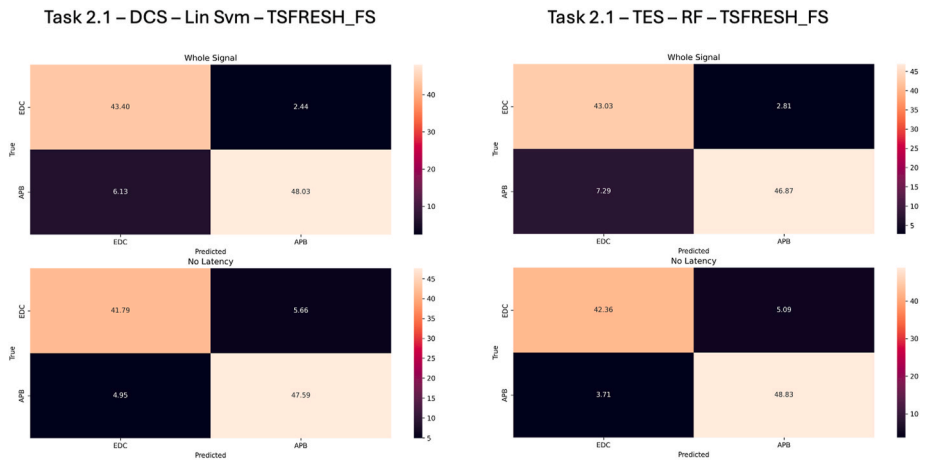

Fig. 5. Confusion matrices for Task 2.1 On the left, the confusion matrices for task 2.1 on DCS acquisition modality with reference to the best performing model and signal representation. On the right, the confusion matrix for task 2.1 on TES acquisition modality with reference to the best performing model and signal representation.

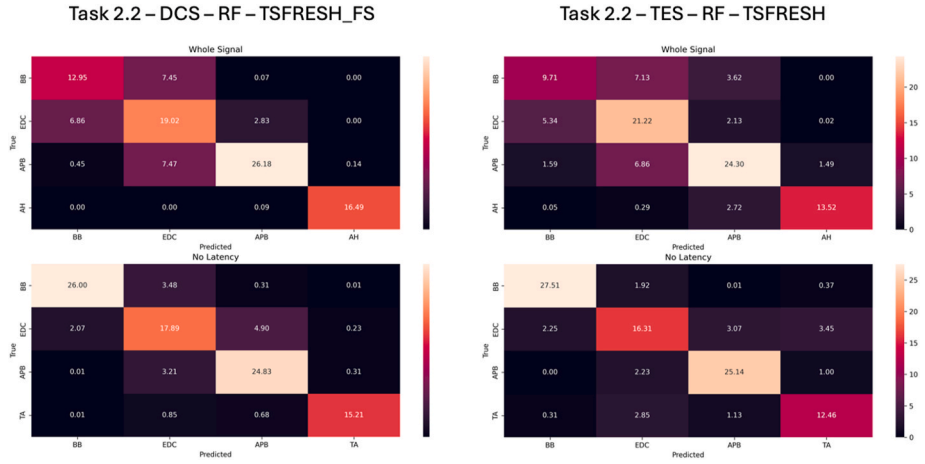

Fig. 6. Confusion matrices for Task 2.2. On the left, the confusion matrices for task 2.2 on DCS acquisition modality with reference to the best performing model and signal representation. On the right, the confusion matrix for task 2.2 on TES acquisition modality with reference to the best performing model and signal representation.

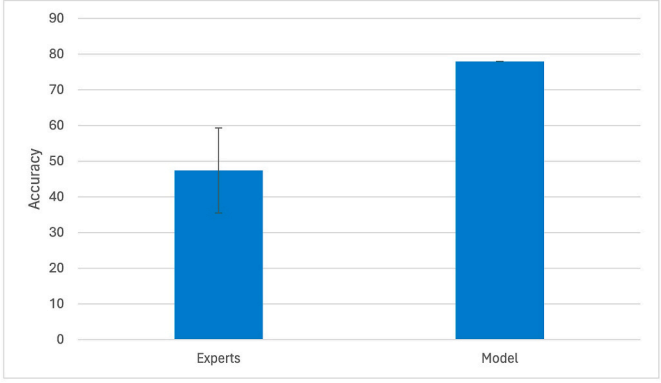

Fig. 7. Accuracy performance comparison between the best performing ML model and ten human experts. The plot shows the classification accuracy performance expressed as the percentage of correctly classified signals, between ten human experts (mean and standard deviation) compared to the one from the best performing model (Random Forest) in task 3.

was prospectively validated.

4 Discussion

We have demonstrated that ML provides efficient analytical techniques to classify intraoperative MEPs coming from the main muscles monitored during a neurosurgical operation to the brain or the spine. Overall, the best performing model proved to be Random Forest applied to MEPs expressed as a set of mathematically derived features (TSFRESH/TSFRESH_FS representation) with 94 % accuracy when considering two muscles from the upper limb (no latency database), 84 % when considering four muscle from upper and lower limbs (whole signal database), and 79 % when considering six muscles from upper and lower limbs and mixed MEP acquisition protocols (whole signal database). Our best performing model proved to be significantly more accurate in identifying the correct muscles from their signals compared to expert neurophysiologists.

The interpretation of biological electrical signals is challenging as they are inherently complex and present a certain degree of variability within and between subjects. The use of Pattern Recognition/Machine Learning and time-series analysis techniques have greatly improved our ability to extract meaningful information from complex signals, such as those found in biomedical domains. These techniques have been used to

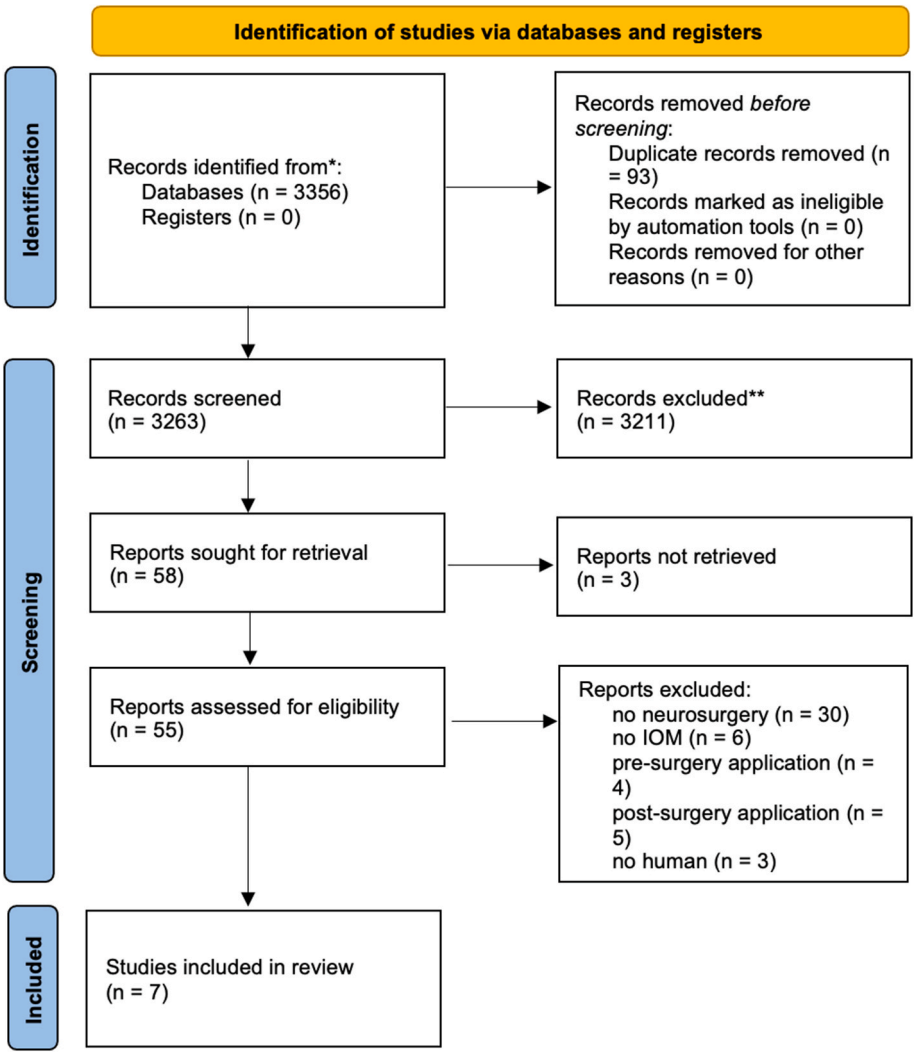

Fig. 8. PRISMA flow diagram of the articles included in the review.

identify patterns in several different signals, from electrocardiograms to fMRI scans, from EEG to EMG, and have led to significant advancements in areas such as disease diagnosis and treatment as well as creation of brain-machine interfaces to vicariate impaired functions [22–24, 46–51].

The novel approach we provided in our work was first focused on evaluating if there was actual information within MEPs that could be extracted by ML models. By conducting model training and testing within the same patient, Task 1 provided convincing evidence about the possibility to extract muscle-specific information, obtaining over 99 % classification accuracy for the best performing model. Such evidence provided the basis to use ML to classify MEPs coming from different subjects that we tested on task 2. Here we could see that, while a two-class classification problem was easier (task 2.1 - best performance 94 %) compared to correctly classifying MEPs from four different muscles (task 2.2 - best performance 84 %), the peak performances remained very high in both cases. It has to be considered that in task 2.1 the models were able to account for a greater level of variability as more patients were included compared to task 2.2; in addition, the two muscles (apb and edcb) included in task 2.1 were by far the most represented in terms of number of signals. We can, therefore, reasonably expect an additional improvement in task 2.2 performance by increasing the number of patients and signals involved.

Finally, we conducted task 3 in order to try and take advantage of the information provided by both DCS and TES protocols and in an attempt to further generalize our model by considering all the six muscle classes traditionally included in an intraoperative MEP monitoring setting. In this case as well, the models were able to extract and retain useful information with a peak performance of 79 %.

It was interesting to see how the importance of latency, a key element used by trained neurophysiologists to evaluate an MEP, was limited in determining high classification performances in most of our models. In fact, little difference is observed between the whole signal and no latency databases both in the two muscle and in the TES four muscle task performances. Interestingly, we observed that the models tended to do slightly better by not considering the latency when trying to classify transcranial MEPs coming from two muscle located in the same limb (Table 5). Nevertheless, latency still retains its informative value in the most difficult task where six different muscles located in upper and lower limbs are considered, allowing an improvement in performance of 13 % when considering the MEP in its entirety (Fig. 9).

It is important to do some considerations with regards to the differences in performance observed over the spectrum of ML models also in relation to the different signal representations involved. Random Forest proved to be the overall best classifier due to its ability to handle high dimensional and complex data.

Random Forest is an ensemble method that combines multiple decision trees to make a final prediction. The decision trees in a Random Forest are different, being built on different random subsets of objects and/or features. The final prediction is made by averaging the output of

Table 7
Summary of studies applying machine learning approaches to IOM data in neurosurgery.

Author	Type of IOM technique involved	Additional data streams considered	Population characteristics	Surgery	Type of ML approach	Model aim	Variables included in the model	Model performance	Interpretability assessment	Prospective validation	Comparison to human/ traditional methods performance
N. Qiao et al. [41]	Flash VEP monitoring	No	76 patients with sellar tumors	transphenoidal tumor resection	Three layer convolutional neural network, pretrained convolutional neural network, combination of a convolutional and recurrent neural network	detect no VEP change, min 25 % increase, or 25 % decrease compared to baseline	amplitude from the positive P2 peak at around 120 ms to the preceding N2 negative peak at around 90 ms	accuracy range 82.7–87.4 % sensitivity range 78.9–92.6 % specificity range 80.5–100 %	Class Activation Map	No	No
D.J. Caldwell et al. [43]	ECoG arrays and DBS lead recordings	No	patients with medically intractable epilepsy	invasive electrophysiological mapping and DBS	unsupervised hierarchical density-based clustering technique	define dictionary of artifacts templates	raw multi-channel traces	no ground truth, visual spectral overlap between filtering methods	Spectral features overlap	No	Yes
P. Asman et al. [42]	SSEP	No	8 patients with primary brain tumors of the Rolandic area	craniotomy for tumor resection	unsupervised spectral clustering	central sulcus delineation	raw or derivative SSEP traces	accuracy 85.8–96.3 %	Heat Map	No	Yes
M. Kim et al. [39]	EMG	No	50 patients with hemifacial spasm due to vascular compression	microvascular decompression	convolutional neural network	detect presence or absence of lateral spread response	EMG screenshot images	AUC 0.96	Heat Map	No	Yes
M.R. Jamaludin et al. [44]	TcMEP	No	55 patients with lumbar spine disease	decompression/ instrumentation/ correction surgery	KNN, bagged trees	predict positive functional outcome	onset latency, peak-to-peak amplitude in μV, AUC	sensitivity 75%–100 %, specificity 0–33.3 %	No	No	Yes
S. Rampp et al. [40]	EMG	Yes	200 patients with vestibular schwannoma	vestibular schwannoma resection surgery	Neural Network	estimate post-operative facial nerve deficit	traintime, tumor size, pre-op facial nerve function	Chi squared 51.3, CI 49.7–53.0 - Cramer's V 0.36	No	No	Yes
J. Wermelinger et al. [45]	TcMEP	No	36 patients with brain tumor or vascular pathology	tumor resection, vascular pathology exclusion	RF, KNN, LogReg	TcMEP muscle classification	pre-processed raw signals	two muscles accuracy 43–97 %, four muscles accuracy 28–83 %	No	No	Yes
Current study	TcMEP/ DcMEP	No	54 patients with brain or spinal cord tumor	tumor resection surgery	KNN, linSVM, rbfSVM, RF	MEP muscle classification within a single patiens, between different patients and between different iom settings (tcMEP vs dcMEP)	tcMEP and dcMEP expressed as raw signals or as a set of clinically relevant/ mathematically extracted features	single patient accuracy 49–99 %, inter-patient accuracy 45–93 %, inter iom protocol accuracy 44–79 %	No	No	Yes

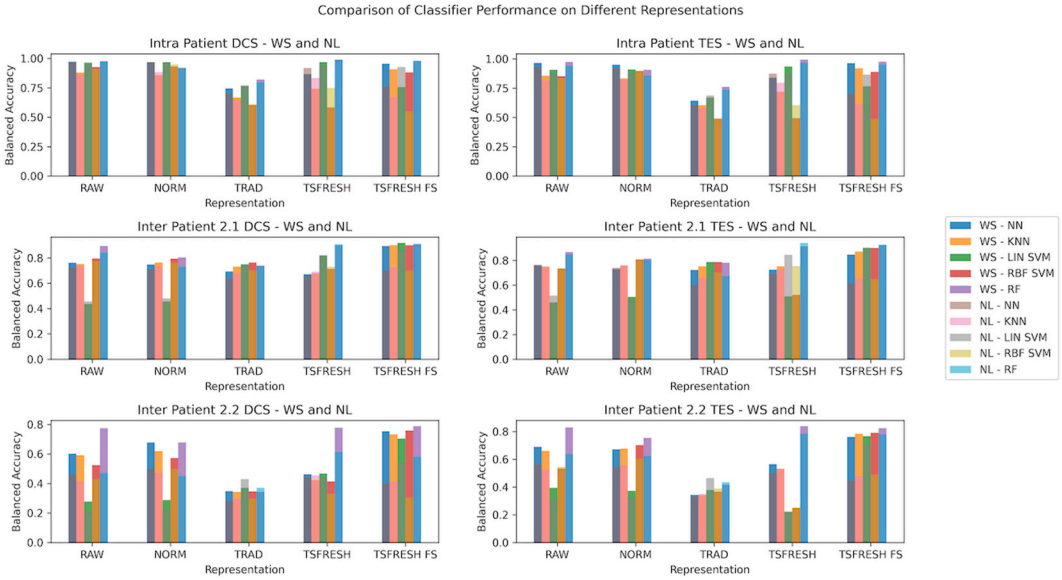

Fig. 9. Bar chart of models' performances. In this bar chart, we plotted the balanced accuracy results of every pair classifier-representation for tasks 1, 2.1 and 2.2 on both whole signal and no latency databases. WS: whole signal, NL: no latency.

all the trees, thus reducing overfitting, which is a common problem in the classification of medical data, where the data can be highly variable and complex [52–54]

With regards to the signal representation considered, all the models performed worse, also compared to the human experts, when trained on traditional MEP features (TRAD) while excelled when considering the signals in their raw format or as features obtained from mathematical formulas whose meaning is not immediately apparent even to the trained eye of human experts. Such findings suggest that ML models work differently from the human mind and, while the information provided by some more traditional MEP features might still be there - as is depicted by the differences in performance between the whole signal database vs the no latency database -, they probably appear in a different form, as these models have very efficient but different ways of extracting meaningful information. Also, it is important to observe that feature selection permits to reduce the impact of the *curse of dimensionality* – a set of problems which may appear if the number of features is too large with respect to the number of objects in the training set [55]. By selecting the most informative features compared to considering the raw signal in its entirety, we can extract a more compact representation which allows a better generalization to unseen patterns [53].

The highly systematic and analytic approach of Machine Learning models is probably the reason why they proved to be significantly better at classifying signals in comparison to human experts; it has to be noted, to be fair, that experts in intraoperative neurophysiology are not trained in recognizing different muscles but rather on evaluating changes over time compared to MEP baseline. Such a scenario presents an important opportunity for a direct clinical application of our model which, by reliably matching an MEP to its muscle, it could be used as a clinical decision support system. In the preparation phase of an IOM setting in fact, instead of relying solely on the expertise of the neurophysiologist, such a tool would confirm the correctness of the connections between the patients' actual muscles to the labels manually given before starting the surgery, therefore preventing muscle mislabeling and erroneous connections that could potentially lead to catastrophic clinical consequences [56,57].

Despite the apparent clinical relevance of potential applications of ML in IOM, there is a dearth of literature with regards to signal analysis and ML application in the realm of intraoperative neurophysiology. In the review work we conducted we found only seven papers focused on the application of ML to IOM data in neurosurgery. Most of them aimed at providing a prognostic tool to estimate or predict post-surgical outcomes. Examples are Jamaludin et al. who explored the possibility to predict positive functional outcomes of patients undergoing lumbar spine surgery using three traditional MEP features (onset latency, peak-to-peak amplitude in μV, and AUC) to train and test different ML models with peak sensitivity and specificity achieved by Fine kNN of 87.5 % and 33.33 %, respectively [44]. Qiao et al. tested different deep learning models to classify visual evoked potentials (VEPs) changes during transsphenoidal surgery with good differentiation ability between no change, improvement and reduction, comparable to human experts [41]; and Rampp et al. combined EMG data with pre-operative nerve functionality score and tumor size to estimate post-operative facial nerve deficit [40]. Two works presented results that could have a direct impact in the adjustment of an actual surgical strategy: Asman et al. successfully applied an unsupervised clustering algorithm to SSEPs heat maps assessed with ECoG grids to delineate the central sulcus automatically with high precision, without the need for peak and latency tracking [42], and Caldwell et al. provided a dictionary of signal artifacts that could be used to refine the IOM signal in real time, facilitating its interpretation [43].

Finally, a recent work by Wermelinger et al. explored the potentialities of standard ML models in classifying pre-processed and raw transcranial MEPs on a population of 36 patients with good preliminary results in classifying MEPs from two muscles (89 % same limb and 97 % different limbs) and four muscles (83 %).

Our work significantly advances the preliminary works reported mainly along two directions. On one hand, we better investigated the effect of the representation of the neurophysiological signal in the ML pipeline; in Qiao et al., Kim et al. and Wermelinger et al., the authors mainly focused on the classification part, leaving the representation to standard choices; in our work we studied more in depth this crucial issue, showing that a careful set of advanced signal processing features (TSFRESH and TSFRESH_FS) permits to better characterize the MEP, improving the classification accuracy and increasing explainability, while reducing the computational load. As a second point, we investigated the potentialities of the proposed approach within different scenarios (called tasks in our paper): rather than simply quantifying how easy it is to classify a muscle on a set on previously unseen subjects, we also studied the discrimination capabilities of the approach within the same subject, and across different protocols (TES and DCS). Furthermore, for all these scenarios, we considered both the whole signal and

the signal without latency, this last scenario permitting to assess the possibility of discriminating the muscles on the sole basis of the MEP waveform, which would allow a more general application of our model also in clinical conditions that affect latency but not necessarily the signal waveform.

It was interesting to see that, while most of the models did not present completely explainable inner dynamics, there was almost consistently an attempt to provide a form of model interpretation using heat and activation maps. Similarly, almost all the papers provided comparison between the models' performance and traditional methods or human performance at the same task (Table 7).

The next step in light of additional future clinical application of such algorithms, consists in the clinical characterization of the features involved as well as the exploration of the correlation between signal changes and different clinical states. With the continued advancement of technology and the increasing availability of large-scale data sets, we can expect to see even more breakthroughs in the field of pattern recognition and time-series analysis of biomedical data. All these considerations highlight the importance of collaborative work between healthcare professionals and data scientists, as it is of fundamental importance to choose the right model and the right data representation for each specific problem [8,32,58,59]

Some important limitations have to be considered in the interpretation of our results. Most importantly, the limited number of patients along with the use of a single institution database and one type of monitoring machine, limits the generalizability of our models. Secondly, whilst we considered the overall classification accuracy in each task, the differences in the number of signals available for each muscle might make the models ablity to recognize some muscles better than others. Finally, in this work we included only signals that were considered 'normal', so the classification performance we found cannot be considered representative of pathological MEPs.

5 Conclusions

In this initial era of exploration of ML in IOM, we found that carefully selected and trained ML models, have the ability to extract meaningful information to identify patient muscles from their intraoperative MEPs using a limited number of mathematical features. In this regard, Random Forest proved to be robust across patients and signal acquisition modalities, with the capacity of outperforming human experts and with the potential to act as decision support system to the IOM team.

Such encouraging findings lay the path to further explore the underlying nature of clinically important signals, with the aim to continue to produce useful applications to be used to make surgeries safer and more efficient.

CRediT authorship contribution statement

Alessandro Boaro: Writing – review & editing, Writing – original draft, Supervision, Data curation, Conceptualization. **Alberto Azzari:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Federica Basaldella:** Writing – review & editing, Methodology, Data curation. **Sonia Nunes:** Writing – review & editing, Validation, Data curation. **Alberto Feletti:** Writing – review & editing, Investigation. **Manuele Bicego:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Francesco Sala:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that there are no known conflicts of interest associated with this publication and there has been no significa this work that could have influenced its outcome. The manuscript has been read and approved by all named authors. The study was approved by the institutional Review Board.

Aknowledgments

We would like to thank the precious work of the experts in intraoperative neurophysiology Andrea Badari, Fabrizio Baldanzi, Federica Basaldella, Alessandro Borio, Laura Caldana, Cristiana Martinelli, Sonia Nunes, Giovanna Maddalena Squintani, Giulia Masi, Sara Rinaldo, Vincenzo Tramontano and Simone Troiano who provided their expertise for the expert-based classification of MEPs.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109032.

References

[1] V. Deletis, F. Sala, Intraoperative neurophysiological monitoring of the spinal cord during spinal cord and spine surgery: a review focus on the corticospinal tracts, Clin. Neurophysiol. 119 (2) (2008) 248–264, https://doi.org/10.1016/j.clinph.2007.09.135.

[2] F. Sala, A. Bricolo, F. Faccioli, P. Lanteri, M. Gerosa, Surgery for intramedullary spinal cord tumors: the role of intraoperative (neurophysiological) monitoring, Eur. Spine J. 16 (S2) (2007) 130–139, https://doi.org/10.1007/s00586-007-0423-x.

[3] F. Sala, P. Manganotti, V. Tramontano, A. Bricolo, M. Gerosa, Monitoring of motor pathways during brain stem surgery: what we have achieved and what we still miss? Neurophysiol Clin Neurophysiol 37 (6) (2007) 399–406, https://doi.org/10.1016/j.neucli.2007.09.013.

[4] W.J. Levy, Transcranial stimulation of the motor cortex to produce motor-evoked potentials, Med. Instrum. 21 (5) (1987) 248–254.

[5] S. Tsutsui, H. Yamada, Basic Principles and recent Trends of transcranial motor evoked potentials in intraoperative neurophysiologic monitoring, Neurol. Med.-Chir. 56 (8) (2016) 451–456, https://doi.org/10.2176/nmc.ra.2015-0307.

[6] T. Kombos, O. Süss, Neurophysiological basis of direct cortical stimulation and applied neuroanatomy of the motor cortex: a review, Neurosurg. Focus 27 (4) (2009) E3, https://doi.org/10.3171/2009.8.FOCUS09141.

[7] F. Ringel, F. Sala, Intraoperative mapping and monitoring in supratentorial tumor surgery, J. Neurosurg. Sci. 59 (2) (2015) 129–139.

[8] A. Boaro, M. Mammi, O. Arnaout, Chapter 75: artificial intelligence and Big data in neurosurgery. Youmans & Winn neurological surgery, in: Youmans & Winn Neurological Surgery, eighth ed., 2023.

[9] A. Boaro, F. Sala, Intraoperative neurophysiology during intramedullary spinal cord tumor surgery, in: C.N. Seubert, J.R. Balzer, Sloan Koht (Eds.), Toleikis's Monitoring the Nervous System for Anesthesiologists and Other Health Care Professionals, Springer International Publishing, 2023, pp. 635–645, https://doi.org/10.1007/978-3-031-09719-5_34.

[10] E. Asimakidou, P.A. Abut, A. Raabe, K. Seidel, Motor evoked potential warning criteria in supratentorial surgery: a Scoping review, Cancers 13 (11) (2021) 2803, https://doi.org/10.3390/cancers13112803.

[11] D.B. MacDonald, Motor evoked potential warning criteria, J. Clin. Neurophysiol. 34 (1) (2017) 1–3, https://doi.org/10.1097/WNP.0000000000000346.

[12] R.N. Holdefer, D.B. MacDonald, S.A. Skinner, Somatosensory and motor evoked potentials as biomarkers for post-operative neurological status, Clin. Neurophysiol. 126 (5) (2015) 857–865, https://doi.org/10.1016/j.clinph.2014.11.009.

[13] A. Feletti, A. Boaro, D. Giampiccolo, G. Casoli, F. Moscolo, M. Ferrara, F. Sala, et al., Spinal hemangioblastomas: analysis of surgical outcome and prognostic factors, Neurosurg. Rev. 45 (2) (2022) 1645–1661, https://doi.org/10.1007/s10143-021-01696-x.

[14] S.M. Goetz, B. Luber, S.H. Lisanby, A.V. Peterchev, A novel model Incorporating two variability Sources for describing motor evoked potentials, Brain Stimulat 7 (4) (2014) 541–552, https://doi.org/10.1016/j.brs.2014.03.002.

[15] S.M. Goetz, S.M.M. Alavi, Z.D. Deng, A.V. Peterchev, Statistical model of motor-evoked potentials, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (8) (2019) 1539–1545, https://doi.org/10.1109/TNSRE.2019.2926543.

[16] K. Machetanz, A.L. Gallotti, M.T. Leao Tatagiba, M. Liebsch, L. Trakolis, S. Wang, et al., Time-frequency representation of motor evoked potentials in brain tumor patients, Front. Neurol. 11 (2021), https://doi.org/10.3389/fneur.2020.633224.

[17] J.T. Senders, P.C. Staples, A.V. Karhade, M.M. Zaki, W.B. Gormley, M.L. D. Broekman, et al., Machine learning and neurosurgical outcome prediction: a systematic review, World Neurosurg 109 (2018) 476–486.e1, https://doi.org/10.1016/j.wneu.2017.09.149.

[18] A. Boaro, J.R. Kaczmarzyk, V.K. Kavouridis, M. Harary, M. Mammi, H. Dawood, et al., Deep neural networks allow expert-level brain meningioma segmentation and present potential for improvement of clinical practice, Sci. Rep. 12 (1) (2022 Sep 14) 15462, https://doi.org/10.1038/s41598-022-19356-5.

[19] C.M.W. Goedmakers, A.M. Lak, A.H. Duey, A.W. Senko, O. Arnaout, M.W. Groff, et al., Deep learning for Adjacent segment disease at Preoperative MRI for Cervical Radiculopathy, Radiology 301 (3) (2021 Dec) 664–671, https://doi.org/10.1148/radiol.2021204731.

[20] C. Lega, M. Pirruccio, M. Bicego, L. Parmigiani, L. Chelazzi, L. Cattaneo, The Topography of visually guided Grasping in the Premotor cortex: a Dense-transcranial Magnetic stimulation (TMS) mapping study, J. Neurosci. 40 (35) (2020) 6790–6800, https://doi.org/10.1523/JNEUROSCI.0560-20.2020.

[21] A. Boaro, J. Leung, H.T. Reeder, F. Siddi, E. Mezzalira, G. Liu, et al., Smartphone GPS signatures of patients undergoing spine surgery correlate with mobility and current gold standard outcome measures, J. Neurosurg. Spine 35 (6) (2021) 796–806, https://doi.org/10.3171/2021.2.SPINE202181.

[22] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, et al., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat Med 25 (1) (2019) 65–69, https://doi.org/10.1038/s41591-018-0268-3.

[23] F. Di Nardo, A. Nocera, A. Cucchiarelli, S. Fioretti, C. Morbidoni, Machine learning for detection of muscular activity from surface EMG signals, Sensors 22 (9) (2022) 3393, https://doi.org/10.3390/s22093393.

[24] L.R. Quitadamo, F. Cavrini, L. Sbernini, F. Riillo, L. Bianchi, S. Seri, et al., Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: a review, J. Neural. Eng. 14 (1) (2017) 011001, https://doi.org/10.1088/1741-2552/14/1/011001.

[25] E. Stålberg, H. Erdem, Quantitative motor unit potential analysis in routine, Electromyogr Clin Neurophysiol 42 (7) (2002) 433–442.

[26] M. Christ, N. Braun, J. Neuffer, A.W. Kempa-Liehr, Time series FeatuRe extraction on basis of scalable Hypothesis tests (tsfresh – a Python package), Neurocomputing 307 (2018) 72–77, https://doi.org/10.1016/j.neucom.2018.03.067.

[27] Li Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, et al., Feature selection: a data Perspective, ACM Comput. Surv. 50 (6) (2018) 1–45, https://doi.org/10.1145/3136625.

[28] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[29] B.F. Darst, K.C. Malecki, C.D. Engelman, Using recursive feature elimination in random forest to account for correlated variables in high dimensional data, BMC Genet. 19 (Suppl 1) (2018) 65, https://doi.org/10.1186/s12863-018-0633-8.

[30] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, S. Homayouni, Support vector machine versus random forest for Remote sensing image classification: a Meta-analysis and systematic review, IEEE J Sel Top Appl Earth Obs Remote Sens 13 (2020) 6308–6325, https://doi.org/10.1109/JSTARS.2020.3026724.

[31] P. Cunningham, S.J. Delany, K-nearest Neighbour classifiers - a Tutorial, ACM Comput. Surv. 54 (6) (2022) 1–25, https://doi.org/10.1145/3459665.

[32] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[33] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.

[34] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the Stratification of Multi-label data, in: D. Gunopulos, T. Hofmann, D. Malerba, M. Vazirgiannis (Eds.), Machine Learning and Knowledge Discovery in Databases, vol. 6913, Springer, Berlin Heidelberg, 2011, pp. 145–158, https://doi.org/10.1007/978-3-642-23808-6_10. Lecture Notes in Computer Science.

[35] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000 May) 412–424, https://doi.org/10.1093/bioinformatics/16.5.412. PMID: 10871264.

[36] D. Chicco, G. Jurman, The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification, BioData Min. 16 (1) (2023 Feb 17) 4, https://doi.org/10.1186/s13040-023-00322-4. PMID: 36800973; PMCID: PMC9938573.

[37] D.J. Hand, R.J. Till, A Simple Generalisation of the area under the ROC Curve for multiple class classification problems, Mach. Learn. 45 (2001) 171–186, https://doi.org/10.1023/A:1010920819831.

[38] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71. Available online: https://www.bmj.com/content/372/bmj.n71.

[39] M. Kim, S.K. Park, Y. Kubota, S. Lee, K. Park, D.S. Kong, Applying a deep convolutional neural network to monitor the lateral spread response during microvascular surgery for hemifacial spasm, PLoS One 17 (11) (2022 Nov 2) e0276378, https://doi.org/10.1371/journal.pone.0276378. PMID: 36322573; PMCID: PMC9629649.

[40] S. Rampp, M. Holze, C. Scheller, C. Strauss, J. Prell, Neural networks for estimation of facial palsy after vestibular schwannoma surgery, J. Clin. Monit. Comput. 37 (2) (2023 Apr) 575–583, https://doi.org/10.1007/s10877-022-00928-9. Epub 2022 Nov 4. Erratum in: J Clin Monit Comput. 2022 Dec 8;: PMID: 36333576; PMCID: PMC10068649.

[41] N. Qiao, M. Song, Z. Ye, W. He, Z. Ma, Y. Wang, et al., Deep learning for automatically visual evoked potential classification during surgical Decompression of Sellar region tumors, Transl Vis Sci Technol 8 (6) (2019) 21, https://doi.org/10.1167/tvst.8.6.21.

[42] P. Asman, S. Prabhu, D. Bastos, S. Tummala, S. Bhavsar, T.M. McHugh, et al., Unsupervised machine learning can delineate central sulcus by using the spatiotemporal characteristic of somatosensory evoked potentials, J. Neural. Eng. 18 (4) (2021) 046038, https://doi.org/10.1088/1741-2552/abf68a.

[43] D.J. Caldwell, J.A. Cronin, R.P.N. Rao, K.L. Collins, K.E. Weaver, A.L. Ko, et al., Signal recovery from stimulation artifacts in intracranial recordings with dictionary learning, J. Neural. Eng. 17 (2020) 026023.

[44] M.R. Jamaludin, K.W. Lai, J.H. Chuah, M.A. Zaki, K. Hasikin, N.A. Abd Razak, et al., Machine learning application of transcranial motor-evoked potential to predict positive functional outcomes of patients, Comput. Intell. Neurosci. 2022 (2022) 1–13, https://doi.org/10.1155/2022/2801663.

[45] J. Wermelinger, Q. Parduzi, M. Sariyar, A. Raabe, U.C. Schneider, K. Seidel, Opportunities and challenges of supervised machine learning for the classification of motor evoked potentials according to muscles, BMC Med Inform Decis Mak 23 (1) (2023 Oct 2) 198, https://doi.org/10.1186/s12911-023-02276-3. PMID: 37784044; PMCID: PMC10544622.

[46] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358, https://doi.org/10.1056/NEJMra1814259.

[47] J.P. Martínez, R. Almeida, S. Olmos, A.P. Rocha, P. Laguna, A wavelet-based ECG delineator: evaluation on standard databases, IEEE Trans. Biomed. Eng. 51 (4) (2004) 570–581, https://doi.org/10.1109/TBME.2003.821031.

[48] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. 32 (3) (1985) 230–236, https://doi.org/10.1109/TBME.1985.325532. BME-.

[49] I. Ahmad, X. Wang, M. Zhu, C. Wang, Y. Pi, J.A. Khan, et al., EEG-based Epileptic seizure detection via machine/deep learning approaches: a systematic review, Comput. Intell. Neurosci. 2022 (2022) 1–20, https://doi.org/10.1155/2022/6486570.

[50] K.B. Mikkelsen, J.K. Ebajemito, M.A. Bonmati-Carrion, N. Santhi, V.L. Revell, G. Atzori, et al., Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy, J. Sleep Res. 28 (2) (2019), https://doi.org/10.1111/jsr.12786.

[51] Ascent of machine, Ascent of machine learning in medicine, Nat. Mater. 18 (5) (2019) 407, https://doi.org/10.1038/s41563-019-0360-1, 407.

[52] E. Alickovic, A. Subasi, Medical decision support system for diagnosis of Heart arrhythmia using DWT and random forests classifier, J. Med. Syst. 40 (4) (2016) 108, https://doi.org/10.1007/s10916-016-0467-8.

[53] K. Li, N. Yu, P. Li, S. Song, Y. Wu, Y. Li, et al., Multi-label spacecraft electrical signal classification method based on DBN and random forest, PLoS One 12 (5) (2017) e0176614, https://doi.org/10.1371/journal.pone.0176614.

[54] D. Petry, C. Mirian de Godoy Marques, J.L. Brum Marques, Baroreflex sensitivity with different lags and random forests for staging cardiovascular autonomic neuropathy in subjects with diabetes, Comput. Biol. Med. 127 (2020 Dec) 104098, https://doi.org/10.1016/j.compbiomed.2020.104098. Epub 2020 Oct 28. PMID: 33152669.

[55] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[56] H.N. Modi, S.W. Suh, J.H. Yang, J.Y. Yoon, False-negative transcranial motor-evoked potentials during scoliosis surgery causing paralysis: a case report with literature review, Spine 34 (24) (2009 Nov 15) E896–E900, https://doi.org/10.1097/BRS.0b013e3181b40d4f. PMID: 19910760.

[57] C.D. Yingling, Are there false-negative and false-positive motor-evoked potentials? J. Clin. Neurophysiol. 28 (6) (2011 Dec) 607–610, https://doi.org/10.1097/WNP.0b013e31823db022. PMID: 22146357.

[58] D.A. Hashimoto, G. Rosman, D. Rus, O.R. Meireles, Artificial intelligence in surgery: Promises and Perils, Ann. Surg. 268 (1) (2018) 70–76, https://doi.org/10.1097/SLA.0000000000002693.

[59] D. Denisko, M.M. Hoffman, Classification and interaction in random forests, Proc Natl Acad Sci 115 (8) (2018) 1690–1692, https://doi.org/10.1073/pnas.1800256115.