Neurosurgical Review
https://doi.org/10.1007/s10143-019-01163-8

REVIEW

Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review

Quinlan D. Buchlak[1] · Nazanin Esmaili[1,2] · Jean-Christophe Leveque[3,4] · Farrokh Farrokhi[3,4] · Christine Bennett[1] · Massimo Piccardi[5] · Rajiv K. Sethi[3,4,6]

Received: 22 May 2019 / Revised: 5 July 2019 / Accepted: 6 August 2019

Abstract

Machine learning (ML) involves algorithms learning patterns in large, complex datasets to predict and classify. Algorithms include neural networks (NN), logistic regression (LR), and support vector machines (SVM). ML may generate substantial improvements in neurosurgery. This systematic review assessed the current state of neurosurgical ML applications and the performance of algorithms applied. Our systematic search strategy yielded 6866 results, 70 of which met inclusion criteria. Performance statistics analyzed included area under the receiver operating characteristics curve (AUC), accuracy, sensitivity, and specificity. Natural language processing (NLP) was used to model topics across the corpus and to identify keywords within surgical subspecialties. ML applications were heterogeneous. The densest cluster of studies focused on preoperative evaluation, planning, and outcome prediction in spine surgery. The main algorithms applied were NN, LR, and SVM. Input and output features varied widely and were listed to facilitate future research. The accuracy ($F_{(2,19)} = 6.56$, $p < 0.01$) and specificity ($F_{(2,16)} = 5.57$, $p < 0.01$) of NN, LR, and SVM differed significantly. NN algorithms demonstrated significantly higher accuracy than LR. SVM demonstrated significantly higher specificity than LR. We found no significant difference between NN, LR, and SVM AUC and sensitivity. NLP topic modeling reached maximum coherence at seven topics, which were defined by modeling approach, surgery type, and pathology themes. Keywords captured research foci within surgical domains. ML technology accurately predicts outcomes and facilitates clinical decision-making in neurosurgery. NNs frequently outperformed other algorithms on supervised learning tasks. This study identified gaps in the literature and opportunities for future neurosurgical ML research.

Keywords Artificial intelligence · Deep brain stimulation · Deep learning · Machine learning · Neurosurgery · Risk stratification · Spine surgery

Introduction

Artificial intelligence (AI) involves the use of computer systems to achieve goals by simulating cognitive capabilities [42, 79]. Machine learning (ML) classification is a domain of AI that enables algorithms, or classifiers, to learn patterns in large, complex datasets and generate useful predictive outputs [55, 78, 87]. The application of ML algorithms to new large datasets can reveal novel trends and relationships that may have beneficial implications for clinical practice in medicine [23]. Multiple reviews have investigated the application of ML methods in healthcare and have demonstrated the substantial impact of ML in generating improvements to healthcare quality and safety [63, 69, 71]. Neurosurgery in particular is a high-risk field continually seeking to minimize complications and improve surgical techniques and outcomes. The modern

✉ Quinlan D. Buchlak
quinlan.buchlak1@my.nd.edu.au

1 School of Medicine, The University of Notre Dame, Sydney, NSW, Australia

2 Rozetta Institute, Sydney, NSW, Australia

3 Neuroscience Institute, Virginia Mason Medical Center, Seattle, WA, USA

4 Department of Neurosurgery, Virginia Mason Medical Center, Seattle, WA, USA

5 University of Technology Sydney, Sydney, NSW, Australia

6 Department of Health Services, University of Washington, Seattle, WA, USA

Published online: 17 August 2019

neurosurgeon works in a technological environment with access to many data inputs before, during, and after surgery [16], and ML has the potential to be used in new ways to improve the safety of neurosurgery for patients and increase the likelihood of positive outcomes.

ML can be broadly split into four main types of learning: supervised, unsupervised, reinforcement, or semi-supervised [79, 97, 111]. Supervised learning trains algorithms with datasets that contain pre-labeled outcomes for each case in order to solve classification and regression problems, while unsupervised learning uses unlabeled input data and allows the algorithm to extract features and patterns on its own, which is often known as clustering. Succinctly, supervised learning uses input variables (X) to predict a defined outcome (Y), whereas unsupervised learning involves the use of input variables (X) to extract features, patterns, and structure from the data. Reinforcement learning uses the principles of behaviorism (reward and punishment) to generate the algorithm program [110]. Many applications use a semi-supervised learning pattern, in which only a subset of the data is labeled with an output variable of interest, as fully labeled data can be expensive to acquire [121].

Supervised and unsupervised methods require the selection and application of appropriate algorithms to generate meaningful conclusions from data [87]. Supervised learning algorithms include logistic regression (LR), support vector machines (SVM), neural networks (NN), deep neural networks (DNN), decision trees (DT), random forests (RF), and naïve Bayes (NB) approaches. LR models are often the algorithm of choice for predicting dichotomous outcomes [44]. SVMs are flexible in representing complex relationships but are susceptible to overfitting [91]. NNs appear to perform well as predictive tools partly because they are able to accurately model complex nonlinear relationships in high-volume datasets. A NN is made up of multiple neurons, which share information via weighted connections. Each neuron consists of an activation function that defines its output. Training the network and optimizing its weighted connections involves feeding information through it and feeding information back again, while monitoring a loss function, in processes called forward- and back-propagation. Weights are adjusted until adequate predictions are achieved. NNs are often able to achieve predictive performance that is beyond the capabilities of linear models [1, 44, 83, 87]. Training a NN requires a selection of the best parameters to optimize the predictive performance of the model without causing overfitting [87], which can limit predictive generalisability to novel cases. DNNs are an expanded version of standard NNs, incorporating many more hidden layers and neurons. Both NN and DNN algorithms have seen substantial success in modern ML research, particularly in medical image analysis. To optimize performance, any statistical model and ML approach, including NNs, endeavors to find the lowest point in a mathematical error surface. The shape of this surface is dictated by the combination of the training data and the complexity of the model. DNNs are by far the most complex models in current use, with numbers of free parameters in the order of tens or hundreds of millions. To obtain smooth error surfaces that can be effectively minimized, they correspondingly require orders of magnitude more data than other algorithms (e.g., LR) [20, 43, 56]. In return, DNNs have consistently proved able to deliver much higher predictive performance than conventional models. DT algorithms use a set of supervised learned decision rules to make predictions for the target variable, and, while they are relatively efficient computationally and easier to interpret, they often lack generalization to novel cases [62, 86]. RFs are an aggregation of DTs and determine outputs by combining the predictions of individual trees. They resist overfitting and mitigate the issue of poor generalization faced by DTs [22].

Once an appropriate algorithm has been selected, a predictive model is developed through appropriate training, calibration, validation, and peer review. The performance of supervised ML models is typically evaluated by statistical outputs, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the area under the receiver operating characteristics curve (AUC), among others [36, 49, 76]. These statistics allow the comparison of the performance of different forms of supervised learning models and allow the researcher to determine the applicability of the model to clinical practice [36, 98].

The number of published ML studies in neurosurgery is increasing exponentially [97]. This systematic review was conducted to assess the current state of ML applications to neurosurgery and to gain insight into recent developments, focusing on the application of ML algorithms to support clinical decision-making in neurosurgery across the neurosurgical process continuum (pre-, intra-, and postoperative). We were guided by the following research question: How have ML algorithms been applied to support clinical decision-making in neurosurgery and do NNs demonstrate higher performance than other ML algorithms? Our analytical framework (Fig. 1) guided this analysis.

Method

Our method was guided by the standards of the Institute of Medicine [75] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [74]. A prospective systematic review protocol was developed and approved by senior authors. We developed a comprehensive search strategy [93] using the following databases: PubMed, Medline, Embase, and Scopus to identify all relevant studies up to March 2019.

Search strategy

The search was conducted in February and March 2019, by inserting the keywords “machine learning neurosurgery” into

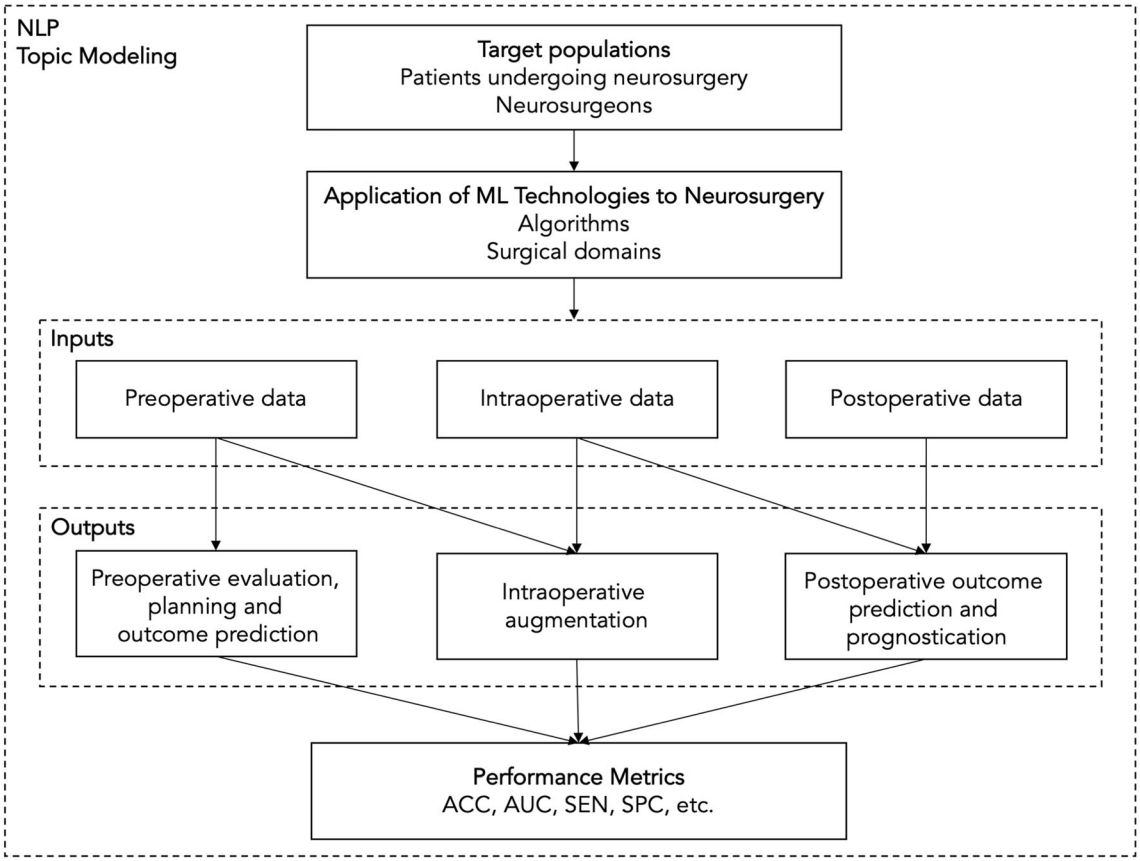

Fig. 1 The analytical framework guiding the analysis underlying this systematic review

PubMed, which generated 351 articles. Inserting this term into Embase and Scopus yielded 1679 articles and 100 articles, respectively. In Medline, combining the MeSH terms Machine Learning and Neurosurgical Procedures using the AND Boolean operator yielded just four results. We inserted the following search term into PubMed ((machine learning OR artificial intelligence OR prediction) AND (neurosurgery OR spine surgery OR neurological surgery OR neurosurgical procedures)) and generated 4883 results. We investigated MeSH terms to identify other relevant keywords and hand searched both referenced and citing articles to identify additional relevant papers [90]. We reviewed the table of contents of well-respected neurosurgical journals to ensure that no relevant papers had been missed by this search protocol. These journals were selected based on impact factor, SCImago Journal Rank indicator and author judgment. Our search strategy employed comprehensive combinations of two categories of search terms: (1) domain-specific and (2) methodological (Table 1).

Inclusion criteria and study selection process

Articles for inclusion had to be English-language, original, peer-reviewed research that included data relevant to neurosurgery describing the application of ML to directly inform and assist neurosurgical clinical decision-making and practice. Studies were excluded if they were conference papers or abstracts or involved non-human subjects. Articles involving the use of ML in medical image analysis without explicit reference or application to neurosurgery were excluded, as the application of ML to medical image analysis has been reviewed elsewhere [50, 65, 97, 104]. Risk of bias was assessed using the Prediction model Risk of Bias Assessment Tool (PROBAST) [118].

After initial title and abstract screening and the removal of 21 duplicates, 129 articles were selected for full-text review. The selection of articles for inclusion was conducted by one researcher (QDB) and verified by a second researcher (NE). Articles classified as appropriately meeting inclusion criteria independently by both researchers were included in the analysis with disagreements resolved by discussion.

Analysis

We conducted a qualitative and quantitative analysis of the studies included. Data extraction involved collecting and coding the following information and variables for each study: abstract, surgical domain, application domain, model inputs, algorithms used, neural network details, algorithm type, model outputs, number of patients, number of algorithms employed, highest performing ML model, and all reported

Table 1 Keywords and MeSH terms used in the search strategy

	Domain-specific terms	Methodological terms
Keywords	• Neurosurgery	• Machine learning
	• Spine surgery	• Neural networks
	• Neurological surgery	• Deep neural networks
	• Intraoperative/perioperative processes	• Convolutional neural networks
	• Postoperative outcomes	• Deep learning
	• Preoperative	• Support vector machine
	• Patient safety	• Random forest
	• Quality	• Predictive modeling
	• Value	• Outcome prediction
	• Clinical Decision Support Systems	• Clinical decision support
		• Logistic regression
MeSH terms	• Neurosurgery	• Machine learning
	• Neurosurgical procedures	• Unsupervised machine learning
	• Clinical decision support systems	• Supervised machine learning
	• Clinical decision supports	• Artificial intelligence
	• Prognosis	• Computational intelligence
	• Classification	

model performance statistics. The quantitative analysis was conducted in Python and was based on comparing algorithm performance metrics across studies. The metrics focused on were AUC, accuracy, sensitivity, and specificity because they were the most widely reported. We aggregated performance metrics across studies and performed one-way ANOVA with post hoc (Tukey) tests. The significance threshold was set at $p < 0.05$.

We developed a categorization taxonomy for the studies reviewed based on surgical procedure domain and domain of application. Surgical procedure domains included (1) brain tumor surgery, (2) deep brain stimulation (DBS) surgery, (3) spine surgery, (4) neurovascular surgery, (5) neurosurgery to treat epilepsy, and (6) neurosurgery (other). We organized application domains around the pre-, intra-, and postoperative surgical process continuum [101]. Application domains included (1) preoperative outcome prediction (used preoperative data to predict postoperative outcomes); (2) intraoperative augmentation; and (3) postoperative outcome prediction (used intra- or postoperative data to predict postoperative outcomes). Intraoperative augmentation referred to applications of ML models to assist a neurosurgeon with decision-making during surgery.

Natural language processing (NLP) is a computational method and application of ML used to convert language into a formal representation that can be readily analyzed by computers [68, 87]. NLP analysis involved the use of the Python nltk [18], gensim [88, 89], and pyLDAvis [29, 106] packages. The NLP analysis consisted of two phases: (1) keyword identification by surgical domain and (2) topic modeling using the latent Dirichlet allocation (LDA) approach [19]. For keyword identification, we split abstracts into surgical domain corpora. For each domain, we tokenized the text, converted tokens to lower case, dropped numeric characters, removed English stop words, lemmatized the remaining tokens, and counted the most frequently occurring meaningful words. Topic modeling involved the same pre-processing steps using all abstracts, with the addition of token vectorization and filtering to remove common and rare words. Tokens that appeared in less than 5 abstracts and those that appeared in more than 50% of abstracts were filtered out. A corpus was created, term frequency–inverse document frequency (tf-idf) calculations were conducted and various LDA multicore models were developed to classify up to 30 topics. LDA model performance was assessed with perplexity and coherence metrics and visualization was performed with pyLDAvis.

Results

We found that 70 articles fulfilled criteria for inclusion (Fig. 2). Previous reviews of ML in neurosurgery have focused on providing a general overview [99] comparing ML algorithm performance with that of human experts [97] and the use of ML for outcome prediction [98]. Our systematic review consisted of a substantial corpus of 38 recent neurosurgery ML articles not referenced by these previous reviews. We present an analysis of the algorithms applied, the inputs they were trained on, the outputs they were trained to predict, and their relative performance statistics. The median number of patients in each study was 120 (mean = 1693, SD = 4737).

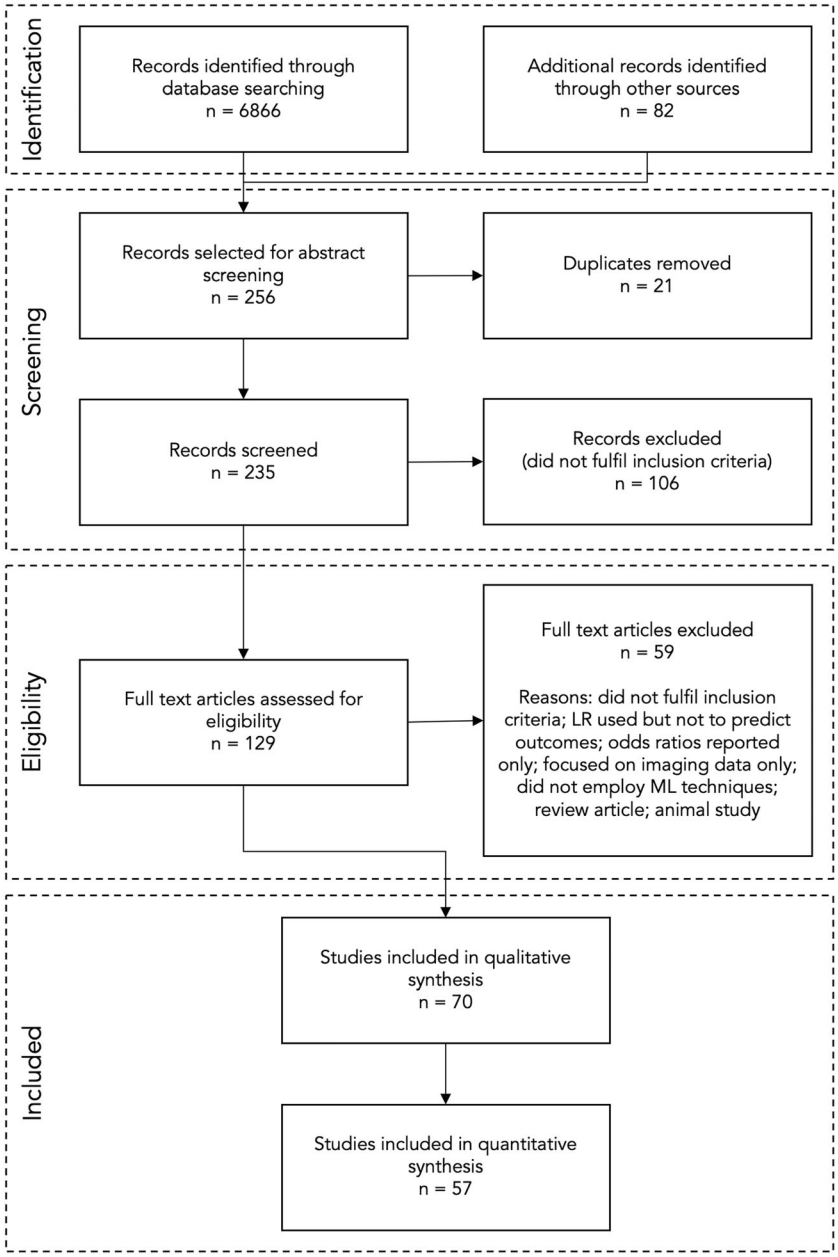

Fig. 2 PRISMA flow diagram of the article selection process

The median number of ML algorithms employed in each study was two (mean = 3.5, SD = 4.8).

Publications and algorithms applied in neurosurgery over time

The number of publications applying ML to neurosurgical decision support has increased rapidly over the past decade (Fig. 3). The top three most frequently applied algorithms were NN, LR, and SVM. Most algorithms appear to have been applied recently in neurosurgery.

Summary of literature included, and algorithms applied

We found substantial heterogeneity in modeling approaches. A range of different types and various combinations of ML algorithms were applied in the studies reviewed (Table 2). LR and NN (40% and 35% of studies) were the most frequently applied algorithms. SVMs were applied in 29% of studies. RF and NB were each applied in 8% and DTs were applied in 7% of the studies reviewed. Hidden Markov models were applied in 4% of studies. Unsupervised learning methods were applied in 6% of studies.

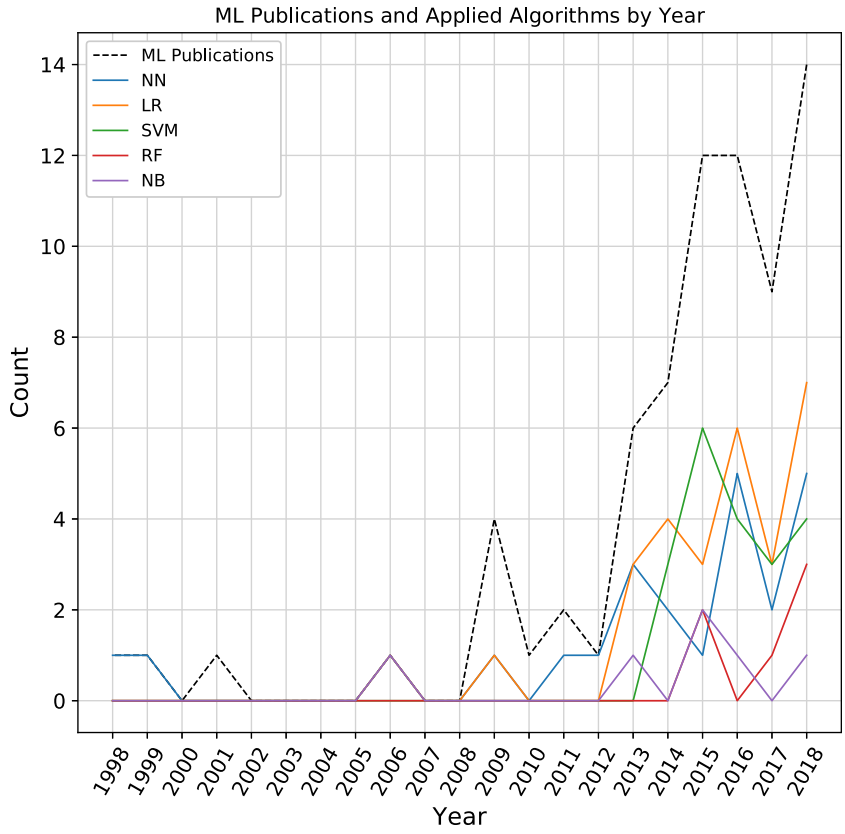

Fig. 3 Count of total ML publications and the number of publications employing specific algorithms by year. For clarity, rarely applied algorithms (KNN, HMM, DT, etc.) are not shown

We developed a synthesis of the variables used as inputs. In predictive modeling research, questions regarding relevant variables to be included often arise. These decisions are guided by variables that are available in an organization's electronic medical record, theoretical importance, and clinician judgment. The synthesis presented provides a useful reference for future ML practitioners in neurosurgery.

Application and surgical domain matrix

Categorizing studies by surgical and application domains revealed clusters of application foci and areas that may benefit from additional research (Table 3). The main clusters of studies focused on preoperative evaluation, planning, and outcome prediction in spine surgery ($n = 12$) and neurosurgery to treat epilepsy ($n = 8$). Overall, spine surgery saw the highest number of ML publications ($n = 19$), followed by brain tumor surgery ($n = 17$). Few ML studies have focused on neurovascular surgery ($n = 4$). Of the application domains, the greatest number focused on preoperative evaluation, planning, and outcome prediction.

Predictive model performance evaluation statistics

There was substantial variation in the way authors chose to evaluate the performance of their models. Accuracy and AUC were the most frequently reported performance statistics (57% and 50% of studies). Sensitivity and specificity were also commonly reported (39% and 36% of studies). Twenty-six percent of studies reported PPV while NPV was reported by 22%. Seven percent of studies reported an *f* statistic and 6% used forms of mean squared error. The Hosmer-Lemeshow test was reported in 6% of studies. Few studies used kappa value [14], Bland-Altman plots [48], coefficients of determination [48], mean absolute difference [48], Brier score [57], Matthews correlation coefficient [62], and reliability [116].

Comparing algorithm performance

A subset of studies compared the performance of different algorithms on the same dataset (Table 4). The most common comparison was between NN and LR. NN outperformed LR (difference in AUC) in 93% of these studies. Three studies compared NN and SVM. NN outperformed SVM in two of these studies. Three studies compared RF and LR. RF outperformed LR in these studies. Three studies compared SVM and NB. SVM demonstrated superior performance. NN and DT were compared twice and NN demonstrated superior performance in both studies. NNs showed superior performance in 86% of studies ($n = 21$) where they were compared with other algorithms.

The performance of the top three most frequently applied algorithms (LR, NN, and SVM) was compared statistically. There was not enough reported performance data in this

Table 2 Original research studies that applied machine learning methods to support decision making in neurosurgery and met inclusion criteria

Reference	Year	Input features	Algorithms applied	Outcomes predicted	Number of patients	Risk of bias
Brain tumor surgery						
Akbari [3]	2016	Demographics	SVM	Early recurrence of glioblastoma	65	+
Akbari [2]	2014	Medical history and examination	SVM, PCA	Highly infiltrated brain tissue	79	?
Azimi [12]	2015	• Neurologic status • Smoking history and current smoking status • Comorbidities	NN, LR	Probability of new cerebral metastases after surgery	192	?
Emblem, Pinho [35]	2014		SVM	Glioma survival at 6 months, 1, 2, and 3 years	235	+
Emblem, Due-Tonnessen [34]	2014	• Hyponatremia • Seizure disorder	SVM	1-, 2-, 3-, and 4-year survival	94	+
Izadyyazdanabadi [51]	2018	• Coagulopathy	NN	Intraoperative tumor diagnosis	74	+
Jermyn [53]	2015	• Post-surgery treatment	DT	Intraoperative glioma detection	17	+
Ji [54]	2015	Investigation data • Tumor histology, subtype and size	GAM	Intraoperative tumor differentiation	18	?
Ling [64]	2018	• MRI (blood volume distribution, resting state fMRI, task fMRI)	LR	Postoperative facial nerve function	106	+
Macyszyn [67]	2016	• Diffusion tensor imaging	SVM	Survival at 6 months	105	-
Missios [72]	2015	Intraoperative data • Surgical path/plan • Stimulated Raman scattering microscopy	LR	Postoperative complications in patients undergoing craniotomies for glioma resection	21,384	+
Oermann [80]	2013	• Facial motor evoked potentials • Parametric motion-based and force-based features	NN, LR	1-year survival in patients with brain metastases treated with radiosurgery	196	+
Panesar [82]	2019	• Geometry of anatomy • Gantry and table angles	NN, LR, SVM, DT	2-year mortality in glioma patients	76	+
Qian [85]	2013	Health service characteristics • Region	KM	Path planning in image-guided neurosurgery	120	?
Shamir [103]	2015	• Location (urban teaching, Urban nonteaching, rural) • Size (small, medium, large)	SVM, NB, RF	Motor outcomes, speech, tremor, rigidity, bradykinesia and akinesia at 1 year. Stimulation and drug dose recommendations.	10	?
Skrobala [107]	2014		NN	Beam and table angles	539	?
Tonutti [115]	2017		NN, SVM	Tumor deformation	6600 data points	-
Valsky [116]	2017		SVM	Real-time electrophysiological detection of ventral STN border	81	-
Vergun [117]	2018		SVM, NB, DT	Postoperative mortality at 18 months	62	+
Wong [119]	2009		UL	Anatomic boundaries of subcortical structures	27	?
DBS surgery						
Angeles [4]	2017	Demographics Medical	SVM, DT, KNN	Rigidity, bradykinesia, tremor	7	-
Baumgarten [14]	2016	history and examination • Comorbidities/rigidity/	NN	Occurrence of pyramidal tract side effects	10	+
Buchlak [26]	2018	bradykinesia/tremor • Medications Investigation data	LR	High- or low-acuity discharge disposition status	135	+
Kostoglou [62]	2017	• Location of DBS electrodes • MRI and CT scans Intraoperative data	RF	Motor improvement of Parkinson's disease patients	20	?
Taghva [113]	2011	• Microelectrode recordings	HMM	Identification of brain location	Simulation	-
Taghva [112]	2010	• Amount of current delivered	HMM	Identification of brain location	Simulation	-
Zaidel [120]	2009		HMM	Real-time STN entry and exit and structural borders	21	?
Spine surgery						
Assi [8]	2014	Demographics	SVM	Postoperative trunk 3D shape	141	-
Azimi [13]	2016	Medical history and examination • Comorbidities	NN, LR	Success of lumbar disc herniation surgery	203	+
Azimi [11]	2015	• Duration of symptoms • Leg/back pain and numbness	NN, LR	Recurrent disc herniation and surgical success at 1 year	402	+
Azimi, Benzel [10]	2014	• ODI	NN, LR	2-year surgical satisfaction	168	+
Bekelis [15]	2014	• Zung Depression Scale	LR		13,660	+

Table 2 (continued)

Reference	Year	Input features	Algorithms applied	Outcomes predicted	Number of patients	Risk of bias
		• Japanese Orthopaedic Association Score • Neurogenic Claudication Outcome Score • Numeric rating of back pain • Duration of symptoms • Walking distance		Myocardial infarction, death, infection, DVT, PE, UTI, length of stay ≥ 3 days, return to OR, stroke		
Buchlak [25]	2017	• Recurrent lumbar disc herniation • Level and type of herniation • Sports activities	LR	Postoperative complications and readmission within 30 days of surgery	136	+
Fan [37]	2016	• Occupational profile • Fine motor function	SVM	Detect SEP abnormalities during surgery	Trained on SEP signals	+
Garcia-Cano [40]	2018	• Age at diagnosis • Prior surgery	RF	Evolution of the shape of the spine	150	+
Hoffman [48]	2015	• Medications	LR, SVM	Postoperative ODI	20	+
Karhade [57]	2018	• Baseline PRO scores • Smoking history	NN, SVM, DT, BP	5-year survival after spinopelvic chordoma surgery	265	+
Khor [58]	2018	• Alcohol and drug use	LR	Leg pain, back pain and ODI	1583	+
Kim, Arvind et al. [59]	2018	Investigation data • Preoperative spine curves	NN, LR	VTE, cardiac and wound complications, mortality	4073	+
Kim, Merrill et al. [60]	2018	• Spinal stenosis ratio • 3D spine models	NN, LR	VTE, cardiac and wound complications, mortality	22,629	+
Konar [61]	2016	• Tumor size and location	LR	Mortality at 6, 12, and 18 months	128	+
Lubelski [66]	2014	• Histologic tumor subtype	LR	C5 nerve root palsy	98	?
Ryu [92]	2018	• Primary tumor site • Extraneural metastasis • Multiple lesions	LR, RF	5-, 7-, and 10-year overall survival in patients with spinal ependymoma	2822	+
Scheer [95]	2017	• Bulging disc • spondylolisthesis Intraoperative data	DT	Complications within 6-weeks (intraoperative and perioperative)	557	+
Shamim [102]	2009	• Somatosensory evoked potentials Postoperative data • Extent and location of surgery	FIS	Failure to improve after lumbar disc surgery at 6-month follow up	501	+
Staartjes [109]	2018	• Surgical radiation use • Adjuvant therapy • Gross total resection	NN, LR	Leg pain, back pain, and ODI	422	+
Neurosurgery to treat epilepsy						
Arle [5]	1999	Demographics	NN	Postoperative seizure control	80	+
Armañanzas [6]	2013	Medical history and examination • Seizure type, frequency and laterality	LR, NB, KNN, EMC	Recovery from epilepsy and surgery	23	+
Bernhardt [17]	2015	• Age at seizure onset	UL	Epilepsy outcomes	114	+
Cohen [30]	2016	• Febrile seizures • Operative and clinical variables • Medication use	SVM, NB	Identify candidates for surgical intervention for drug-resistant epilepsy	200	+
Dian [31]	2015	• Wada examination • Handedness Investigation data	SVM	Identification of seizure onset zones and Engel scores	6	?
Gazit [41]	2016	• MRI/CT/EEG variables • Seizure zones and laterality	LR	Assessment of language lateralization	76	+
Grigsby [44]	1998	• WAIS-R and WAIS-III • Wechsler Memory Scale	NN, DFA	Cessation of seizures (Engel's Class 1 or 2)	87	+
Memarian [70]	2015	• Neurophysiological variables	SVM, NB, DFA	Cessation of seizures	20	+
Munsell [76]	2015	• Rorschach test	SVM	Cessation of seizures	118	+
Taylor [114]	2018	• Linguistic unigrams and bigrams	SVM	Postoperative seizure outcomes	53	+
Yankam Njiwa [77]	2015	• Diagnostic and surgical procedures • Intracarotid amobarbital testing Intraoperative data • Electrocorticography evaluations Postoperative data • Pathology of the tissue resected • Surgical site • Engel outcome scale • Elapsed time since seizure • Postoperative MRI	RF	Cessation of seizures	60	+

Table 2 (continued)

Reference	Year	Input features	Algorithms applied	Outcomes predicted	Number of patients	Risk of bias
		• Periventricular fludeoxyglucose white matter increases				
Neurovascular surgery						
Asadi [7]	2016	Demographics Medical history and examination	NN, SVM, DT	Clinical outcomes at final follow up	199	+
Dumont [32]	2016	• Clinical presentation • GCS	NN	The occurrence of symptomatic cerebral vasospasm	25	+
Dumont [33]	2011	Investigation data • Radiographic factors	NN	The occurrence of symptomatic cerebral vasospasm	113	+
Oermann [81]	2016	• Subarachnoid clot thickness • Aneurysm location and diameter • Transcranial Doppler elevation • Prior embolization • Radiation dose/isocenters Postoperative data • Surgical ligation/embolization • Complications	LR	Favorable outcome at 2 years	1674	+
Neurosurgery (other)						
Abouzari [1]	2009	Demographics Medical history and examination	NN, LR	Glasgow outcome score at discharge	300	+
Azimi, Mohammadi [9]	2014	• Birth weight/prematurity • Previous shunt/revision	NN, LR	ETV success at 6 months	168	+
Campillo-Gimenez [27]	2013	• Type of hydrocephalus • Choroid plexus cauterization	NLP	Detection of surgical site infections	5010	+
Habibi [45]	2016	• Myelomeningocele	NN, LR	Predict shunt infection	148	+
Hale [46]	2018	• intraventricular hemorrhage • Infections • Mechanism of injury	NN	Risk of developing a clinically relevant traumatic brain injury	12,902	+
Mitchell [73]	2013	• GCS/ASA status • Motion sickness/PONV Investigation data	NN	Identification of eloquent and functional cortical networks	13	?
Peng [84]	2006	• CT scan • Midline shift/intracranial air	NN, LR, NB	Postoperative nausea and vomiting	1086	+
Savin [94]	2018	• Hematoma characteristics • Brain atrophy Postoperative data	LR, RF	Healthcare-associated ventriculitis and meningitis	2286	+
Shi [105]	2013	• Type and length of surgery • Duration of anesthesia • EVD/days with EVD • Craniotomy, CSF leakage • SSI/days with SSI • Urinary catheter/UTI • Medications • Convulsions • Intestinal dysfunction • Purulent sputum • Hemodialysis • Pleural drain/days with drain • Feeding tube/central line • Length of stay Health service characteristics • Hospital and surgeon case volume	NN, LR	In-hospital mortality	16,956	+

"+" indicates low risk of bias/low concern regarding applicability, "-" indicates high risk of bias/high concern regarding applicability, and "?" indicates unclear risk of bias/unclear concern regarding applicability. *BP*, Bayes point machine; *DF*A, discriminant function analysis; *DVT*, deep vein thrombosis; *EEG,* electroencephalography; *EMC*, expectation-maximization clustering; *EVD*, extraventricular drain; *FIS*, fuzzy logic-based fuzzy inference system; *GAM*, generalized additive model; *HMM*, hidden Markov model; *KM*, K-means; *KNN*, K-nearest neighbors; *LR,* logistic regression; *(f)MRI*, (functional) magnetic resonance imaging; *NB* naïve Bayes; *NLP*, natural language processing; *NN,* neural network; *ODI*, Oswestry Disability Index; *OR*, operating room; *PCA*, principal components analysis; *P*E, pulmonary embolism; *PONV*, postoperative nausea and vomiting; *PRO*, patient reported outcome; *RF*, random forest; *SSI,* surgical site infection; *STN*, subthalamic nucleus; *SVM*, support vector machine; *UL*, unsupervised learning; *UTI*, urinary tract infection; *WAIS-R*, Wechsler Adult Intelligence Scale—Revised

Table 3 Numbers of included articles, coded by surgical domain and domain of application

Surgical domain	Application domain		
	Preoperative evaluation, planning, and outcome prediction	Intraoperative augmentation	Postoperative outcome prediction and prognostication
Brain tumor surgery	6	5	6
DBS surgery	2	6	2
Spine surgery	12	1	6
Neurovascular surgery	2	0	2
Neurosurgery to treat epilepsy	8	0	3
Neurosurgery (other forms)	2	0	7
Total	32	12	26

corpus to include other algorithms. We compared algorithms on accuracy, AUC, sensitivity, and specificity (Table 5). LR, NN, and SVM differed significantly in their accuracy performance ($F_{(2,19)} = 6.56$, $p < 0.01$). Post hoc testing showed that NNs were significantly more accurate than LR. The accuracy performance of NNs and SVMs did not differ significantly. The three algorithms did not differ significantly with regard to AUC ($F_{(2,56)} = 1.75$, $p = 0.18$) or sensitivity ($F_{(2,16)} = 2.85$, $p = 0.07$). The three algorithms did differ significantly with regard to specificity ($F_{(2,16)} = 5.57$, $p < 0.01$). Post hoc testing showed that SVM demonstrated significantly higher specificity than LR.

Neural network characteristics

Most studies only employed basic NNs. Frequently, the NNs had three layers: one input layer, one hidden layer (generally with < 10 nodes), and an output layer [13, 33, 80]. The details of 22 NNs were reported, including the number of layers in each NN and the number of nodes in each layer. The median number of layers in the NNs (including input and output layers) was three (mean = 4.4, SD = 4.0). The median number of hidden layers was one (mean = 1.5, SD = 0.7). The median number of nodes in the first hidden layer was eight (mean = 9.4, SD = 6.8) and the median number of nodes in the second hidden layer was seven (mean = 8.6, SD = 3.8). It was rare to find a study in this corpus that employed a more sophisticated deep [76, 109] or convolutional [52] NN.

Predictive model validation

Fifty-seven studies (79%) provided details of a validation process for the models they used. The most common methods involved applying various forms of internal cross-validation and holdout datasets.

ML-driven clinical decision support systems

A small subset of studies deployed their predictive models to production as clinical decision support systems to directly facilitate the decision-making of clinicians [25, 26, 58]. Some others were designed as prototype intraoperative clinical decision support systems [37, 54, 115].

NLP keyword identification and topic modeling

The top 10 most common meaningful words in each surgical domain along with their frequency are shown in Table 6. The keywords captured the research foci of each of the surgical domains. Emergent themes included a focus on the patient, the use of artificial neural networks (ANN), employing a data-driven approach, specific surgical techniques and procedures employed, and common disease states considered.

Table 4 Comparing algorithm performance: counts of studies where algorithm A outperformed algorithm B

Inferior performance (algorithm B)	Superior performance (algorithm A)				
	Neural network	Logistic regression	Support vector machine	Random forest	Total
Neural network	–	1	1	0	2
Logistic regression	13	–	1	3	17
Support vector machine	2	0	–	0	2
Decision tree	2	0	1	0	3
Total	17	1	3	3	

Table 5 Statistical comparisons of reported performance metrics for neural network, logistic regression and support vector machine algorithms

	Performance metrics mean (SD; *n*)											
	Accuracy			AUC			Sensitivity			Specificity		
Neural network (NN)	88.89 (8.89; 19)			0.79 (0.12; 27)			83.28 (21.66; 17)			68.01 (21.66; 19)		
Logistic regression (LR)	76.17 (14.34; 12)			0.75 (0.11; 57)			65.05 (26.72; 11)			56.91 (19.54; 13)		
Support vector machine (SVM)	81.85 (6.27; 18)			0.80 (0.07; 12)			63.83 (28.98; 17)			79.83 (13.98; 17)		
ANOVA	$p < 0.01$			$p = 0.18$			$p = 0.07$			$p < 0.01$		
Tukey post hoc tests	G1	G2	Diff	G1	G2	Diff	G1	G2	Diff	G1	G2	Diff
	LR	NN	TRUE	LR	NN	FALSE	LR	NN	FALSE	LR	NN	FALSE
	LR	SVM	FALSE	LR	SVM	FALSE	LR	SVM	FALSE	LR	SVM	TRUE
	NN	SVM	FALSE	NN	SVM	FALSE	NN	SVM	FALSE	NN	SVM	FALSE

ANOVA, Analysis of variance; *Diff*, reject the null hypothesis; *G1*, Group 1. *G2*, Group 2

We used unsupervised LDA to identify the main topics in the database of abstract text. In this model, a "topic" is a set of words that recur with similar frequencies in multiple documents. For this corpus of abstracts, model coherence was maximized at 7 topics (perplexity = − 6.14), representing the highest performing model (Fig. 4).

The seven topics identified (Table 7) were characterized by the following topic themes: modeling approaches, model construction techniques, model performance evaluation, type of surgery, type of pathology, and preoperative outcome prediction. Mapping topics visually with pyLDAvis [29, 106] showed acceptable topic separation and overfitting minimization (Fig. 5). Topic word weights generated by the LDA model are displayed in Table 7, while topic word frequencies in the corpus for the first two topics are displayed in Fig. 5.

Discussion

This study reviewed the application of ML to support clinical decision-making in neurosurgery. Results suggested that the field of ML in neurosurgery is growing rapidly and that there is substantial scope for further applications of ML technologies to neurosurgical data. Results suggested that ML applications and modeling methods have been diverse and that a wide variety of outcomes have been successfully predicted to facilitate decision-making. The number of ML publications in neurosurgery has increased rapidly over the past seven years. Most neurosurgical ML publications were in the domain of spine surgery, followed by brain tumor surgery. Few ML studies have focused on neurovascular surgery. The densest cluster of studies, when organized by application and surgical domains, focused on preoperative evaluation, planning, and outcome prediction in spine surgery. There was substantial heterogeneity in the modeling approaches applied. Studies varied with regard to algorithms applied, input and output variables used, and methods for assessing predictive model performance. LR, NN, and SVM were the most commonly applied algorithms. It appeared that NNs frequently outperformed most other algorithms on supervised learning tasks. Demographic, clinical history, and investigation-related features were often used to predict postsurgical outcomes.

Table 6 Top 10 keywords within each surgical domain identified with text mining

Brain tumor surgery		DBS surgery		Spine surgery		Neurovascular surgery		Neurosurgery to treat epilepsy		Neurosurgery (other)	
Keyword	Freq.	Keyword	Freq.	Keyword	Freq.	Keyword	Freq.	Keyword	Freq.	Keyword	Freq.
patient	60	STN	27	patient	87	model	26	patient	45	ANN	36
model	44	DBS	23	model	87	patient	17	outcome	34	model	30
tumor	38	model	18	surgery	51	outcome	13	seizure	30	patient	22
machine	29	patient	17	complication	32	prediction	10	epilepsy	26	network	16
brain	26	stimulation	15	spine	30	SCV	9	surgery	23	regression	15
imaging	26	Parkinson	12	prediction	30	predictive	9	network	22	using	15
method	26	disease	11	study	28	ANN	7	using	21	study	13
survival	25	brain	11	surgical	26	machine	6	learning	21	shunt	13
learning	24	data	11	variable	22	learning	6	machine	20	predict	12
analysis	23	recording	11		22	analysis	6		18	ETV	12

ANN, artificial neural network; *DBS*, deep brain stimulation; *ETV*, endoscopic third ventriculostomy; *SCV*, symptomatic cerebral vasospasm; *STN*, subthalamic nucleus

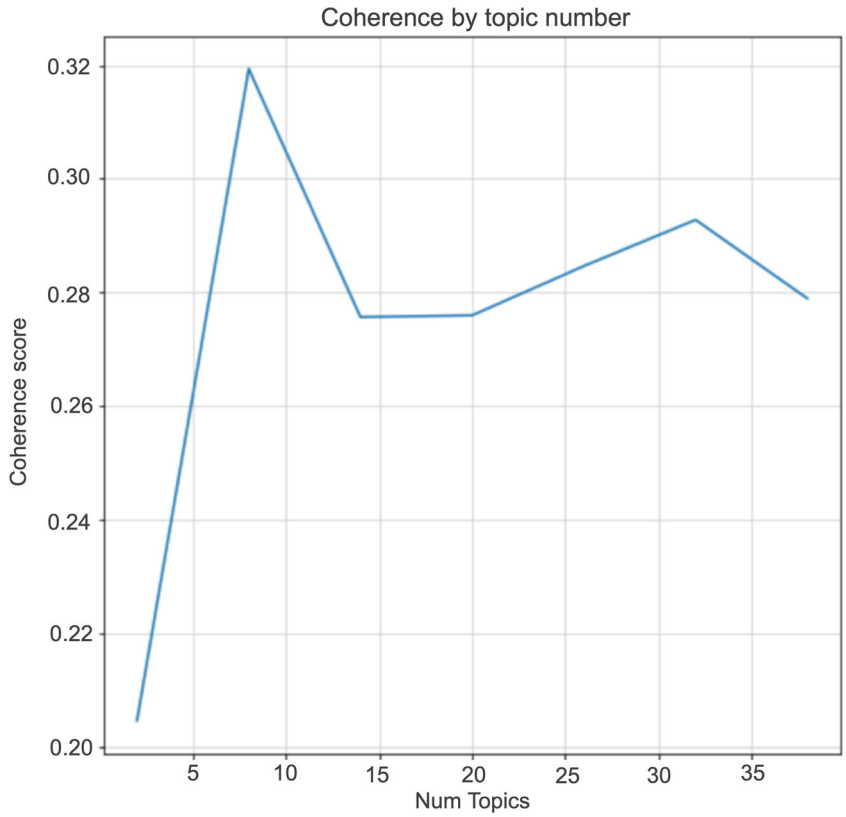

Fig. 4 Coherence plotted against the number of topics. Coherence reached a maximum at 7 topics

Accuracy, AUC, sensitivity, and specificity were the most commonly reported performance metrics. We found that NNs demonstrated significantly higher accuracy than LR and that SVM demonstrated significantly higher specificity than LR. Input and output variables were synthesized to facilitate future modeling work. Gaps and opportunities were identified for clinicians and data scientists to develop additional clinical decision support systems to further improve neurosurgical care.

To our knowledge, this is the first systematic review of ML in neurosurgery to deploy NLP ML methods. By applying NLP technologies, we identified seven distinct topic clusters, and keywords for each of the neurosurgical subdomains. The topics appeared to vary by aspects of the technical modeling approach, patient pathology, and surgical methods. These topics and keywords augment the systematic review by facilitating a deeper, structured understanding of the past foci of the literature and the current state of the science to identify gaps for designing future research projects. Literature gaps are considered in more detail in the section on future research opportunities below.

ML, all the way up from linear regression to deep neural networks, represents a set of powerful technologies capable of effectively predicting outcomes to support decision-making in neurosurgery. It appears that NNs (higher accuracy than LR) and SVMs (higher specificity than LR) have been the top-performing algorithms in the ML armament where large, labeled datasets were available. The main advantage of SVM over LR is its ability to model moderate nonlinearities, whereas LR is limited to linear relations. In turn, NNs can model nonlinearities of, potentially, any complexity. All algorithms, however, are not without their shortcomings [39] and, while algorithms are improving all the time, academic understanding of these shortcomings is still developing. Some algorithms, particularly NNs, are not intuitive, nonlinear, and inscrutable [96] in the way they generate their outputs. The opacity of the mechanics of NNs is a distinct challenge, representing a potential impediment to their widespread adoption. Clinicians tend to lack trust in the outputs of a clinical decision support system when it is not clear how the underlying algorithm came to its classification conclusion. LR allows the clinician to see the weights associated with each variable and provides a sense of how the algorithm classifies cases. This feature is not available in NNs. It is difficult to interpret what they are doing internally and why. Improving the transparency of NNs will be useful in facilitating their implementation and adoption and in achieving associated data-driven improvements in neurosurgery. Future ML researchers should at a minimum report the number of layers and the corresponding number of nodes that comprise the NNs they develop as well as develop better NN visualization methods. Rules have been developed to guide the number of hidden layers and neurons in a neural network, since these model characteristics can affect performance and overfitting [47]. Reporting the number of hidden layers and

Table 7 Topics identified across all article abstracts by the highest performing LDA natural language processing model

Topic	Topic word weights
1	0.013*"threshold" + 0.009*"practice" + 0.009*"time" + 0.008*"system" + 0.008*"better" + 0.008*"real" + 0.008*"mm" + 0.008*"spinal" + 0.007*"error" + 0.007*"current"
2	0.022*"stn" + 0.014*"db" + 0.012*"ann" + 0.010*"author" + 0.008*"microelectrode" + 0.008*"predictive" + 0.008*"infection" + 0.008*"complication" + 0.008*"value" + 0.007*"stimulation"
3	0.015*"ann" + 0.014*"epilepsy" + 0.011*"feature" + 0.010*"performance" + 0.010*"class" + 0.009*"outcome" + 0.009*"seizure" + 0.008*"classifier" + 0.008*"set" + 0.008*"cohort"
4	0.009*"ann" + 0.009*"network" + 0.009*"rate" + 0.009*"group" + 0.008*"spine" + 0.007*"function" + 0.007*"approach" + 0.007*"motor" + 0.007*"brain" + 0.007*"curve"
5	0.010*"ann" + 0.009*"planning" + 0.009*"stimulation" + 0.009*"technique" + 0.008*"also" + 0.008*"decision" + 0.008*"research" + 0.008*"value" + 0.008*"effect" + 0.007*"achieved"
6	0.013*"complication" + 0.011*"spine" + 0.010*"treatment" + 0.010*"postoperative" + 0.009*"function" + 0.009*"improvement" + 0.008*"predicting" + 0.007*"algorithm" + 0.007*"outcome" + 0.007*"p"
7	0.017*"survival" + 0.014*"seizure" + 0.012*"tumor" + 0.011*"imaging" + 0.010*"glioma" + 0.009*"outcome" + 0.008*"tissue" + 0.007*"successful" + 0.007*"association" + 0.007*"symptom"

The weights represent the relative importance of a word to the topic

nodes provides technical insight into the development and operation of the model and facilitates an assessment of its quality. In addition, recently published work on DNNs includes extensive ablation analysis, which consists of removing parts of the network to properly attribute credit for performance and gain insights on its predictions [38].

Safety in neurosurgery is a paramount concern. An issue to be considered in this surgical field is what the consequences may be when a neurosurgeon relies on the guidance of an ML-driven clinical decision support system and the patient goes on to suffer unexpected poor outcomes. A clinical decision support system must only be used by experts and should be just one of many inputs into the decision-making process. The neurosurgeon is the primary actor in a complex network of information processors and advisers, all working together in the service of the patient. To maximize safety, ML models should be developed following a robust process. They should be trained on high-quality datasets, be well validated, and be subjected to peer review. In the US, they may require FDA approval prior to full implementation. Predictive models should be validated on an unseen holdout dataset after they have been trained. This holdout dataset typically consists of a random selection of a proportion (e.g., 20–30%) of the main dataset, which is set aside and not used during the model training phase. Once a predictive model has been developed and validated in this standard way, further external validation, while controlling for operating institution, is likely to be beneficial.

Decision support tools based on trained machine learning algorithms may be integrated into surgical practice in a number of beneficial ways. Like a synthetic subject matter expert, these systems have a role in acting as supporting, complementary, data-driven inputs into a broader multidisciplinary and multimodal surgical risk analysis and mitigation protocol. The predictive risk stratification information they provide should feed into appropriate parts of the surgical decision-making process (e.g., the multidisciplinary conference discussion and/or patient consultation), just as other relevant information currently does (e.g., investigation results and assessments from other specialists). A diagram indicating a viable workflow process integration point for ML decision support tools, within the context of a comprehensive patient review protocol designed for complex spine surgery, has been published by the Seattle Spine Team [24]. A similar process and tool integration point for neurosurgical teams is likely to afford patient safety benefits.

While ML-based systems are powerful technologies, they should not replace the clinical judgment of the physician and the medical team. This technology performs well as an expert assistant on narrow, well-defined tasks. The ideal role of these systems is as data-driven inputs into the surgical decision-making process, which are designed to solve focused problems like predicting the risk of complications for a specified procedure type. Because of the practical complexities of day-to-day surgical practice, they are not yet well suited to broader,

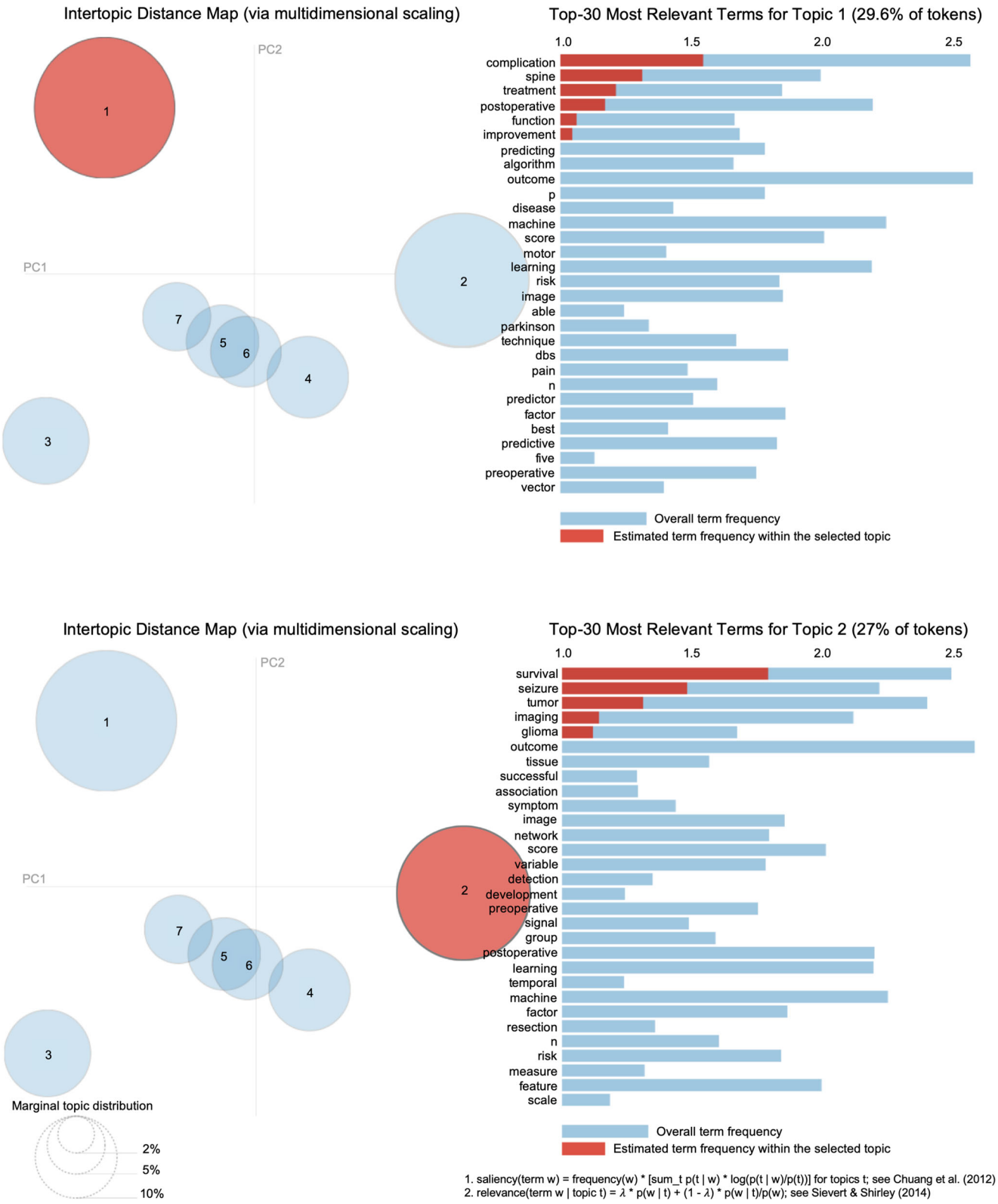

Fig. 5 The first two topics identified by LDA topic modeling, visualized with pyLDAvis [29, 106]

more general medical functions. These systems have a role as assistants to surgeons and, fundamentally, their role is one of augmentation, not automation.

We found that few research groups published details of clinical decision support system tools that applied the models they trained and validated [25, 58]. There is potential for the

development of more application systems (e.g., risk calculators) outside of spine surgery. ML algorithms can only facilitate safety improvements if they are applied as part of a usable system that is adopted by surgeons. If ML is confined to the academic realm, effective risk stratification in complex neurosurgical patients is hindered. Clinical decision support systems that can be widely adopted need to be developed. To that end, groups like the Seattle Spine Team have utilized multidisciplinary clearance conferences to effectively mitigate risk [24, 100]. AI technology may provide ancillary information to clinicians seeking to stratify risk in live patient care.

As the field matures, various types of commercial decision support tools are likely to become available. When considering whether to invest in a commercial tool, surgeons may consider the following three factors to assess its applicability, safety, utility, and quality. The system should (1) be designed to solve the specific problem of concern (e.g., complication risk stratification in DBS surgery); (2) be underpinned by peer-reviewed published literature illustrating its development process, performance and explainability; and (3) adhere to ML model development guidelines and required approvals. These may be useful questions for surgeons to pose to commercial system purveyors to evaluate the quality of their products.

Data from complex neurosurgical procedures is not abundant and complications are rare events leading to our observation of an output class imbalance in many studies. The performance of many ML algorithms can be challenged by the class imbalance issue. This hindrance is a common problem in applying predictive methods to clinical data, particularly for highly specialized surgical procedures and needs to be appropriately addressed in future neurosurgical ML research. Imbalance may generate a bias toward predicting the more prominent outcome state [59]. This bias may be counteracted by distribution over-sampling, sample weighting, synthesis of new data, and prediction post-processing [21].

ML technology has a collection of strengths and weaknesses. The strengths of ML tools include speed and accuracy. These strengths augment the functioning of the clinician. ML tools are well suited to high-velocity and high-volume data processing, bringing to bear all previous institutional experience to positively shape the care of patients in an individualized way. Building ML tools into the neurosurgical workflow may help to empower the surgeon, reduce the likelihood of error, positively engage the patient, and improve surgeon effectiveness. The ability of ML models to provide accurate and individualized prognoses will be useful as healthcare progresses toward a more precise and value-based future. However, ML is not well suited for prediction if the only available datasets are exceedingly noisy, erroneous, obsolete, biased, or small. In these cases, unassisted human judgment may be preferable. ML may not be well suited to predict minor postoperative issues that are subjective to evaluate, difficult to gather high-quality data for, and that may resolve quickly as a patient recovers.

Limitations and future research

Positive and significant results are generally more likely to be published [28, 108]. Publication bias and selective outcome reporting may have influenced our results as we found no studies reporting failed ML modeling activities, although many studies evaluated the performance of multiple algorithms and reported the inferior performance of some.

We found few studies that comprehensively reported predictive model performance statistics. This omission resulted in relatively small sample sizes for our statistical analysis of algorithm performance. Slightly larger sample sizes would likely have demonstrated additional significant performance differences between algorithms. Future studies would do well to report at least AUC, accuracy, sensitivity, and specificity to facilitate robust model performance evaluation and to enable comparison between studies.

It appears that there are many opportunities for future researchers to make a contribution at the confluence of machine learning and neurosurgery. It is clear that there are opportunities to build on the design of studies already published using similar predictors and outcomes in larger datasets and different neurosurgical patient samples to predict similar and additional outcomes at different time points and build deployable systems based on these models. Few practical predictive software systems exist to support the decision-making of neurosurgeons. No systems have been built to predict short-term complications and functional outcomes in neurovascular surgery. No studies have been published on the development of intraoperative augmentation tools to support neurosurgery for epilepsy or neurovascular surgery and only one has been published in spine surgery. Models may be developed to predict the occurrence of neurosurgical pathology (e.g., aneurism, tumor, spinal stenosis) and the trajectory of postoperative patient improvement across subspecialties. Electronic medical record datasets are rich and underutilized resources, ripe for the application of ML. Convolutional neural networks and transfer learning may be applied to video data to facilitate real-time patient examination, diagnosis (e.g., classification of gait disturbance or tremor) and prognosis or intraoperative anatomical identification and surgical navigation. Recurrent neural networks may be used to predict a patient's hour-by-hour postoperative status to facilitate safer proactive patient management. ML-driven systems may be designed to predict intraoperative coagulopathy in real time. There are many opportunities for surgeons to partner with data scientists to continue to improve surgical decision-making and patient safety in neurosurgery.

Conclusion

ML can be used to effectively facilitate decision-making before, during, and after neurosurgery. There was marked

heterogeneity in the application of ML in neurosurgery, which, along with the prevalence of unique studies, suggests that there is substantial potential to further develop research outputs in this field. NNs appeared to outperform other algorithms across neurosurgical subspecialties, accurately predicting a wide range of operative outcomes. Continuing to deploy ML via well-designed clinical decision support systems is likely to further reduce complication rates and improve quality and safety in neurosurgery.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This was a systematic review based on published literature and the PRISMA guidelines and the need for full ethics review was waived by the institution's ethics committee.

Informed consent Not applicable.

References

1. Abouzari M, Rashidi A, Zandi-Toghani M, Behzadi M, Asadollahi M (2009) Chronic subdural hematoma outcome prediction using logistic regression and an artificial neural network. Neurosurg Rev 32:479–484. https://doi.org/10.1007/s10143-009-0215-3
2. Akbari H, Macyszyn L, Da X, Wolf RL, Bilello M, Verma R, O'Rourke DM, Davatzikos C (2014) Pattern analysis of dynamic susceptibility contrast-enhanced MR imaging demonstrates peritumoral tissue heterogeneity. Radiology 273:502–510
3. Akbari H, Macyszyn L, Da X, Bilello M, Wolf RL, Martinez-Lage M, Biros G, Alonso-Basanta M, O'rourke DM, Davatzikos C (2016) Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. Neurosurgery 78:572–580
4. Angeles P, Tai Y, Pavese N, Wilson S, Vaidyanathan R (2017) Automated assessment of symptom severity changes during deep brain stimulation (DBS) therapy for Parkinson's disease. In: 2017 International Conference on Rehabilitation Robotics (ICORR). IEEE, pp 1512–1517
5. Arle JE, Perrine K, Devinsky O, Doyle WK (1999) Neural network analysis of preoperative variables and outcome in epilepsy surgery. J Neurosurg 90:998–1004. https://doi.org/10.3171/jns.1999.90.6.0998
6. Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, Kastanauskaite A, de Sola RG, DeFelipe J, Bielza C, Larrañaga P (2013) Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. PLoS One 8:e62819
7. Asadi H, Kok HK, Looby S, Brennan P, O'Hare A, Thornton J (2016) Outcomes and complications after endovascular treatment of brain arteriovenous malformations: a prognostication attempt using artificial intelligence. World Neurosurg 96:562–569
8. Assi KC, Labelle H, Cheriet F (2014) Statistical model based 3D shape prediction of postoperative trunks for non-invasive scoliosis surgery planning. Comput Biol Med 48:85–93
9. Azimi P, Mohammadi HR (2014) Predicting endoscopic third ventriculostomy success in childhood hydrocephalus: an artificial neural network analysis. J Neurosurg Pediatr 13:426–432
10. Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR (2014) Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis. J Neurosurg Spine 20:300–305
11. Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S (2015) Use of artificial neural networks to predict recurrent lumbar disk herniation. Clin Spine Surg 28:E161–E165
12. Azimi P, Shahzadi S, Sadeghi S (2015) Use of artificial neural networks to predict the probability of developing new cerebral metastases after radiosurgery alone. J Neurosurg Sci
13. Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR (2016) The prediction of successful surgery outcome in lumbar disc herniation based on artificial neural networks. J Neurosurg Sci 60:173–177
14. Baumgarten C, Zhao Y, Sauleau P, Malrain C, Jannin P, Haegelen C (2016) Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation: proof of concept and application to the pyramidal tract side effect induced by pallidal stimulation. J Med Imaging 3:25001
15. Bekelis K, Desai A, Bakhoum SF, Missios S (2014) A predictive model of complications after spine surgery: the National Surgical Quality Improvement Program (NSQIP) 2005-2010. Spine J 14:1247–1255. https://doi.org/10.1016/j.spinee.2013.08.009
16. Bernardo A (2017) The changing face of technologically integrated neurosurgery: today's high-tech operating room. World Neurosurg 106:1001–1014
17. Bernhardt BC, Hong S, Bernasconi A, Bernasconi N (2015) Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics. Ann Neurol 77:436–446
18. Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc, Sebastopol, CA
19. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
20. Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60:223–311
21. Branco P, Torgo L, Ribeiro RP (2016) A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49:31
22. Breiman L (2001) Random forests. Mach Learn 45:5–32
23. Brusko GD, Kolcun JPG, Wang MY (2018) Machine-learning models: the future of predictive analytics in neurosurgery. Neurosurgery 83:E3–E4
24. Buchlak QD, Yanamadala V, Leveque J-C, Sethi R (2016) Complication avoidance with pre-operative screening: insights from the Seattle spine team. Curr Rev Musculoskelet Med 9:316–326. https://doi.org/10.1007/s12178-016-9351-x
25. Buchlak QD, Yanamadala V, Leveque J-C, Edwards A, Nold K, Sethi R (2017) The Seattle spine score: predicting 30-day complication risk in adult spinal deformity surgery. J Clin Neurosci 43:247–255. https://doi.org/10.1016/j.jocn.2017.06.012
26. Buchlak QD, Kowalczyk M, Leveque J-C, Wright A, Farrokhi F (2018) Risk stratification in deep brain stimulation surgery: development of an algorithm to predict patient discharge disposition with 91.9% accuracy. J Clin Neurosci 57:26–32
27. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M (2012) Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform 192:572–575
28. Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG (2004) Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. Jama 291:2457–2465
29. Chuang J, Manning CD, Heer J (2012) Termite: visualization techniques for assessing textual topic models. In: Proceedings of the international working conference on advanced visual interfaces. ACM, pp 74–77

30. Cohen KB, Glass B, Greiner HM, Holland-Bouley K, Standridge S, Arya R, Faist R, Morita D, Mangano F, Connolly B (2016) Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed inform insights 8:BII-S38308
31. Dian JA, Colic S, Chinvarun Y, Carlen PL, Bardakjian BL (2015) Identification of brain regions of interest for epilepsy surgery planning using support vector machines. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp 6590–6593
32. Dumont TM (2016) Prospective assessment of a symptomatic cerebral vasospasm predictive neural network model. World Neurosurg 94:126–130
33. Dumont TM, Rughani AI, Tranmer BI (2011) Prediction of symptomatic cerebral vasospasm after aneurysmal subarachnoid hemorrhage with an artificial neural network: feasibility and comparison with logistic regression models. World Neurosurg 75:57–63
34. Emblem KE, Due-Tonnessen P, Hald JK, Bjornerud A, Pinho MC, Scheie D, Schad LR, Meling TR, Zoellner FG (2014) Machine learning in preoperative glioma MRI: survival associations by perfusion-based support vector machine outperforms traditional MRI. J Magn Reson Imaging 40:47–54
35. Emblem KE, Pinho MC, Zöllner FG, Due-Tonnessen P, Hald JK, Schad LR, Meling TR, Rapalino O, Bjornerud A (2014) A generic support vector machine model for preoperative glioma survival associations. Radiology 275:228–234
36. Esmaili N, Piccardi M, Kruger B, Girosi F (2018) Analysis of healthcare service utilization after transport-related injuries by a mixture of hidden Markov models. PLoS One 13:e0206274
37. Fan B, Li H-X, Hu Y (2016) An intelligent decision system for intraoperative somatosensory evoked potential monitoring. IEEE Trans Neural Syst Rehabil Eng 24:300–307
38. Fawcett C, Hoos HH (2016) Analysing differences between algorithm configurations through ablation. J Heuristics 22:431–458
39. Feng S, Wallace E, Grissom II A, Iyyer M, Rodriguez P, Boyd-Graber J (2018) Pathologies of neural models make interpretations difficult. In: proceedings of the 2018 conference on empirical methods in natural language processing. Pp 3719–3728
40. Garcia-Cano E, Cosío FA, Duong L, Bellefleur C, Roy-Beaudry M, Joncas J, Parent S, Labelle H (2018) Prediction of spinal curve progression in adolescent idiopathic scoliosis using random forest regression. Comput Biol Med 103:34–43
41. Gazit T, Andelman F, Glikmann-Johnston Y, Gonen T, Solski A, Shapira-Lichter I, Ovadia M, Kipervasser S, Neufeld MY, Fried I (2016) Probabilistic machine learning for the evaluation of presurgical language dominance. J Neurosurg 125:481–493
42. Ghahramani Z (2015) Probabilistic machine learning and artificial intelligence. Nature 521:452–459. https://doi.org/10.1038/nature14541
43. Greenspan H, Van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35:1153–1159
44. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Brewster Smith W (1998) Predicting outcome of anterior temporal lobectomy using simulated neural networks. Epilepsia 39:61–66
45. Habibi Z, Ertiaei A, Nikdad MS, Mirmohseni AS, Afarideh M, Heidari V, Saberi H, Rezaei AS, Nejat F (2016) Predicting ventriculoperitoneal shunt infection in children with hydrocephalus using artificial neural network. Childs Nerv Syst 32:2143–2151
46. Hale AT, Stonko DP, Lim J, Guillamondegui OD, Shannon CN, Patel MB (2018) Using an artificial neural network to predict traumatic brain injury. J Neurosurg Pediatr 1:1–8
47. Heaton J (2008) Introduction to neural networks with Java. Heaton Research, Inc, Chesterfield, MO
48. Hoffman H, Lee SI, Garst JH, Lu DS, Li CH, Nagasawa DT, Ghalehsari N, Jahanforouz N, Razaghy M, Espinal M (2015) Use of multivariate linear regression and support vector regression to predict functional outcome after surgery for cervical spondylotic myelopathy. J Clin Neurosci 22:1444–1449
49. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. Wiley
50. Işın A, Direkoğlu C, Şah M (2016) Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci 102:317–324
51. Izadyyazdanabadi M, Belykh E, Mooney M, Eschbacher J, Nakaji P, Yang Y, Preul MC (2018) Prospects for theranostics in neurosurgical technology: empowering confocal laser endomicroscopy diagnostics via deep learning. arXiv Prepr arXiv180409873
52. Izadyyazdanabadi M, Belykh E, Mooney M, Martirosyan N, Eschbacher J, Nakaji P, Preul MC, Yang Y (2018) Convolutional neural networks: ensemble modeling, fine-tuning and unsupervised semantic localization for neurosurgical CLE images. J Vis Commun Image Represent 54:10–20
53. Jermyn M, Mok K, Mercier J, Desroches J, Pichette J, Saint-Arnaud K, Bernstein L, Guiot M-C, Petrecca K, Leblond F (2015) Intraoperative brain cancer detection with Raman spectroscopy in humans. Sci Transl Med 7:274ra19
54. Ji M, Lewis S, Camelo-Piragua S, Ramkissoon SH, Snuderl M, Venneti S, Fisher-Hubbard A, Garrard M, Fu D, Wang AC (2015) Detection of human brain tumor infiltration with quantitative stimulated Raman scattering microscopy. Sci Transl Med 7:309ra163
55. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science (80-) 349:255–260. https://doi.org/10.1126/science.aaa8415
56. Juan-Albarracín J, Fuster-Garcia E, Manjón JV, Robles M, Aparici F, Martí-Bonmatí L, García-Gómez JM (2015) Automated glioblastoma segmentation based on a multiparametric structured unsupervised classification. PLoS One 10:e0125143
57. Karhade AV, Thio Q, Ogink P, Kim J, Lozano-Calderon S, Raskin K, Schwab JH (2018) Development of machine learning algorithms for prediction of 5-year spinal chordoma survival. World Neurosurg 119:e842–e847. https://doi.org/10.1016/j.wneu.2018.07.276
58. Khor S, Lavallee D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, Lu D, Mohit AA, Oskouian RJ, Roh JR (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 153:634–642
59. Kim JS, Arvind V, Oermann EK, Kaji D, Ranson W, Ukogu C, Hussain AK, Caridi J, Cho SK (2018) Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. Spine Deform 6:762–770
60. Kim JS, Merrill RK, Arvind V, Kaji D, Pasik SD, Nwachukwu CC, Vargas L, Osman NS, Oermann EK, Caridi JM (2018) Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine (Phila Pa 1976) 43:853–860
61. Konar SK, Maiti TK, Bir SC, Kalakoti P, Bollam P, Nanda A (2016) Predictive factors determining the overall outcome of primary spinal glioblastoma multiforme: an integrative survival analysis. World Neurosurg 86:341–348
62. Kostoglou K, Michmizos KP, Stathis P, Sakas D, Nikita KS, Mitsis GD (2017) Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings. IEEE Trans Biomed Eng 64:1123–1130
63. Liang Z, Zhang G, Huang JX, Hu QV (2014) Deep learning for healthcare decision making with EMRs. In: bioinformatics and biomedicine (BIBM), 2014 IEEE international conference on. IEEE, pp 556–559
64. Ling M, Tao X, Ma S, Yang X, Liu L, Fan X, Jia G, Qiao H (2018) Predictive value of intraoperative facial motor evoked potentials in

vestibular schwannoma surgery under 2 anesthesia protocols. World Neurosurg 111:e267–e276
65. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
66. Lubelski D, Derakhshan A, Nowacki AS, Wang JC, Steinmetz MP, Benzel EC, Mroz TE (2014) Predicting C5 palsy via the use of preoperative anatomic measurements. Spine J 14:1895–1901. https://doi.org/10.1016/j.spinee.2013.10.038
67. Macyszyn L, Akbari H, Pisapia JM, Da X, Attiah M, Pigrish V, Bi Y, Pal S, Davuluri RV, Roccograndi L, Dahmane N, Martinez-Lage M, Biros G, Wolf RL, Bilello M, O'Rourke DM, Davatzikos C (2016) Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro-Oncology 18:417–425. https://doi.org/10.1093/neuonc/nov127
68. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. Pp 55–60
69. Manogaran G, Lopez D (2017) A survey of big data architectures and machine learning algorithms in healthcare. Int J Biomed Eng Technol 25:182–211
70. Memarian N, Kim S, Dewar S, Engel J Jr, Staba RJ (2015) Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. Comput Biol Med 64:67–78
71. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 19(6):1236–1246
72. Missios S, Kalakoti P, Nanda A, Bekelis K (2015) Craniotomy for glioma resection: a predictive model. World Neurosurg 83:957–964
73. Mitchell TJ, Hacker CD, Breshears JD, Szrama NP, Sharma M, Bundy DT, Pahwa M, Corbetta M, Snyder AZ, Shimony JS (2013) A novel data-driven approach to preoperative mapping of functional cortex using resting-state functional magnetic resonance imaging. Neurosurgery 73:969–983
74. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151:264–269
75. Morton S, Berg A, Levit L, Eden J (2011) Finding what works in health care: standards for systematic reviews. National Academies Press, Washington DC
76. Munsell BC, Wee C-Y, Keller SS, Weber B, Elger C, da Silva LAT, Nesland T, Styner M, Shen D, Bonilha L (2015) Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. Neuroimage 118:219–230
77. Njiwa JY, Gray KR, Costes N, Mauguiere F, Ryvlin P, Hammers A (2015) Advanced [18F] FDG and [11C] flumazenil PET analysis for individual outcome prediction after temporal lobe epilepsy surgery for hippocampal sclerosis. NeuroImage Clin 7:122–131
78. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567
79. Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 375:1216–1219
80. Oermann EK, Kress M-AS, Collins BT, Collins SP, Morris D, Ahalt SC, Ewend MG (2013) Predicting survival in patients with brain metastases treated with radiosurgery using artificial neural networks. Neurosurgery 72:944–952
81. Oermann EK, Rubinsteyn A, Ding D, Mascitelli J, Starke RM, Bederson JB, Kano H, Lunsford LD, Sheehan JP, Hammerbacher J (2016) Using a machine learning approach to predict outcomes after radiosurgery for cerebral arteriovenous malformations. Sci Rep 6:21161
82. Panesar SS, D'Souza RN, Yeh F-C, Fernandez-Miranda JC (2019) Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. World Neurosurg X:100012
83. Patel JL, Goyal RK (2007) Applications of artificial neural networks in medical science. Curr Clin Pharmacol 2:217–226
84. Peng SY, Wu KC, Wang JJ, Chuang JH, Peng SK, Lai YH (2006) Predicting postoperative nausea and vomiting with the application of an artificial neural network. BJA Br J Anaesth 98:60–65
85. Qian Y, Hui R, Gao X (2013) 3D CBIR with sparse coding for image-guided neurosurgery. Signal Process 93:1673–1683
86. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106
87. Raschka S, Mirjalili V (2017) Python machine learning. Packt Publishing Ltd
88. Řehůřek R (2011) Scalability of semantic analysis in natural language processing
89. Rehurek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on new challenges for NLP frameworks. Citeseer
90. Richards D (2008) Handsearching still a valuable element of the systematic review. Evid Based Dent 9:85
91. Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson education limited, Malaysia
92. Ryu SM, Lee S-H, Kim E-S, Eoh W (2018) Predicting survival of spinal ependymoma patients using machine learning algorithms with SEER database. World Neurosurg 124:e331–339
93. Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D (2008) No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 61:748–754
94. Savin I, Ershova K, Kurdyumova N, Ershova O, Khomenko O, Danilov G, Shifrin M, Zelman V (2018) Healthcare-associated ventriculitis and meningitis in a neuro-ICU: incidence and risk factors selected by machine learning approach. J Crit Care 45:95–104
95. Scheer JK, Smith JS, Schwab F, Lafage V, Shaffrey CI, Bess S, Daniels AH, Hart RA, Protopsaltis TS, Mundis GM (2017) Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J Neurosurg Spine 26:736–743
96. Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham L Rev 87:1085
97. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, Smith TR (2017) Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 83(2):181–192
98. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR, Arnaout O (2018) Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg 109:476–486
99. Senders JT, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, Smith TR, Arnaout O (2018) An introduction and overview of machine learning in neurosurgical care. Acta Neurochir 160:29–38
100. Sethi RK, Pong RP, Leveque J-C, Dean TC, Olivar SJ, Rupp SM (2014) The Seattle Spine Team approach to adult deformity surgery: a systems-based approach to perioperative care and subsequent reduction in perioperative complication rates. Spine Deform 2:95–103
101. Sethi RK, Buchlak QD, Leveque J-C, Wright AK, Yanamadala VV (2018) Quality and safety improvement initiatives in complex spine surgery. In: Seminars in Spine Surgery 30(2):111–120
102. Shamim MS, Glasgow M, Neurosurgery F, Enam SA, Ire F, Sn F (2009) Fuzzy Logic in neurosurgery : predicting poor outcomes

after lumbar disk surgery in 501 consecutive patients. Surg Neurol 72:565–572. https://doi.org/10.1016/j.surneu.2009.07.012
103. Shamir RR, Dolber T, Noecker AM, Walter BL, McIntyre CC (2015) Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease. Brain Stimul 8:1025–1032
104. Shen D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
105. Shi H-Y, Hwang S-L, Lee K-T, Lin C-L (2013) In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. J Neurosurg 118:746–752
106. Sievert C, Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore, MD pp 63–70
107. Skrobala A, Malicki J (2014) Beam orientation in stereotactic radiosurgery using an artificial neural network. Radiother Oncol 111:296–300
108. Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, Hing CB, Harvey I (2009) Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol 9:79
109. Staartjes VE, Marlies P, Vandertop WP, Schröder ML (2018) Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar diskectomy: feasibility of center-specific modeling. Spine J 19(5):853–861
110. Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. In: Introduction to reinforcement learning. MIT press Cambridge
111. Suykens JAK (2014) Introduction to machine learning. Academic Press Library in Signal Processing 1:765–773
112. Taghva A (2010) An automated navigation system for deep brain stimulator placement using hidden Markov models. Oper Neurosurg 66:ons-108
113. Taghva A (2011) Hidden semi-Markov models in the computerized decoding of microelectrode recording data for deep brain stimulator placement. World Neurosurg 75:758–763
114. Taylor PN, Sinha N, Wang Y, Vos SB, de Tisi J, Miserocchi A, McEvoy AW, Winston GP, Duncan JS (2018) The impact of epilepsy surgery on the structural connectome and its relation to outcome. NeuroImage Clin 18:202–214
115. Tonutti M, Gras G, Yang G-Z (2017) A machine learning approach for real-time modelling of tissue deformation in image-guided neurosurgery. Artif Intell Med 80:39–47
116. Valsky D, Marmor-Levin O, Deffains M, Eitan R, Blackwell KT, Bergman H, Israel Z (2017) Stop! Border ahead: automatic detection of subthalamic exit during deep brain stimulation surgery. Mov Disord 32:70–79
117. Vergun S, Suhonen JI, Nair VA, Kuo JS, Baskaya MK, Garcia-Ramos C, Meyerand EE, Prabhakaran V (2018) Predicting primary outcomes of brain tumor patients with advanced neuroimaging MRI measures. Interdiscip Neurosurg 13:109–118
118. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 170:51–58
119. Wong S, Baltuch GH, Jaggi JL, Danish SF (2009) Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during DBS surgery with unsupervised machine learning. J Neural Eng 6:26006
120. Zaidel A, Spivak A, Shpigelman L, Bergman H, Israel Z (2009) Delimiting subterritories of the human subthalamic nucleus by means of microelectrode recordings and a Hidden Markov Model. Mov Disord 24:1785–1793
121. Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3:1–130