


# Interpretable Machine Learning–Based Prediction of Intraoperative Cerebrospinal Fluid Leakage in Endoscopic Transsphenoidal Pituitary Surgery: A Pilot Study

Pier Paolo Mattogno<sup>1,\*</sup> Valerio M. Caccavella<sup>1,\*</sup>  Martina Giordano<sup>1</sup> Quintino G. D'Alessandris<sup>1</sup>  
Sabrina Chiloiro<sup>2</sup> Leonardo Tariciotti<sup>3,4</sup> Alessandro Olivi<sup>1</sup> Liverana Lauretti<sup>1</sup>

<sup>1</sup> Department of Neurosurgery, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico Università Cattolica del Sacro Cuore, Rome, Italy

<sup>2</sup> Department of Endocrinology, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico Università Cattolica del Sacro Cuore, Rome, Italy

<sup>3</sup> Unit of Neurosurgery, Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Cà Granda Ospedale Maggiore Policlinico, Milan, Italy

<sup>4</sup> University of Milan, Milan, Italy

Address for correspondence Martina Giordano, MD, Department of Neurosurgery, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico Università Cattolica del Sacro Cuore, Largo Agostino Gemelli, 8 00168 Rome, Italy (e-mails: msmgiordano@gmail.com; marty-gio@live.it).

J Neurol Surg B Skull Base 2022;83:485–495.

## Abstract

**Purpose** Transsphenoidal surgery (TSS) for pituitary adenomas can be complicated by the occurrence of intraoperative cerebrospinal fluid (CSF) leakage (IOL). IOL significantly affects the course of surgery predisposing to the development of postoperative CSF leakage, a major source of morbidity and mortality in the postoperative period. The authors trained and internally validated the Random Forest (RF) prediction model to preoperatively identify patients at high risk for IOL. A locally interpretable model-agnostic explanations (LIME) algorithm is employed to elucidate the main drivers behind each machine learning (ML) model prediction.

**Methods** The data of 210 patients who underwent TSS were collected; first, risk factors for IOL were identified via conventional statistical methods (multivariable logistic regression). Then, the authors trained, optimized, and audited a RF prediction model.

**Results** IOL reported in 45 patients (21.5%). The recursive feature selection algorithm identified the following variables as the most significant determinants of IOL: Knosp's grade, sellar Hardy's grade, suprasellar Hardy's grade, tumor diameter (on X, Y, and Z axes), intercarotid distance, and secreting status (nonfunctioning and growth hormone [GH] secreting). Leveraging the predictive values of these variables, the RF prediction model achieved an area under the curve (AUC) of 0.83 (95% confidence interval [CI]: 0.78; 0.86), significantly outperforming the multivariable logistic regression model (AUC = 0.63).

**Conclusion** A RF model that reliably identifies patients at risk for IOL was successfully trained and internally validated. ML-based prediction models can predict events that were previously judged nearly unpredictable; their deployment in clinical practice may result in improved patient care and reduced postoperative morbidity and healthcare costs.

## Keywords

- ▶ random forest
- ▶ cerebrospinal fluid leak
- ▶ machine learning
- ▶ pituitary adenoma
- ▶ transsphenoidal surgery
- ▶ pituitary surgery

\* These authors contributed equally to the work.

received  
March 12, 2021  
accepted after revision  
November 12, 2021  
published online  
January 16, 2022

© 2022, Thieme. All rights reserved.  
Georg Thieme Verlag KG,  
Rüdigerstraße 14,  
70469 Stuttgart, Germany

DOI <https://doi.org/10.1055/s-0041-1740621>.  
ISSN 2193-6331.

## Introduction

The endoscopic endonasal transsphenoidal approach has been established as the gold standard for surgery of sellar lesions as a result of the high rates of gross total resection (GTR), low postoperative morbidity, and mortality.<sup>1–4</sup>

The occurrence of intraoperative cerebrospinal fluid (CSF) leak may complicate a variable percentage of transsphenoidal surgery (TSS) procedures, with prevalence estimated at 17.4 to 37.5%.<sup>1,4–6</sup> Intraoperative CSF leakages (IOL) significantly alter the course of surgery in terms of closure strategy, additional invasive maneuvers (e.g., inserting lumbar drainage and harvesting autologous abdominal fat), extended operation time and time under general anesthesia, and the consequent patient morbidity. Notably, IOLs can predispose to the development of postoperative CSF leakages which represent a major source of morbidity and mortality in the postoperative setting.<sup>5,7,8</sup>

Predicting which patients would develop IOL optimizes surgical planning, improves patient counseling and care, reducing postoperative morbidity and the associated costs. While several independent risk factors have been identified for IOLs through univariate and multivariate analysis,<sup>1,4,5,9–11</sup> these express the relationship between a single risk factor and the outcome; however, they fail to capture how the outcome is affected by the interaction of multiple risk factors, with limited integrability into prediction models that may communicate the risk for IOL.

With the recent introduction of artificial intelligence and machine learning (ML) in medicine and neurosurgery, new prediction models have been designed for a range of outcomes, often with a greater area under the curve (AUC) being achieved compared with logistic regression and clinical risk scores.<sup>12–14</sup> Different ML algorithms exist, among which the Random Forest (RF) stands out for its ability to capture nonlinear patterns in the data and the applicability to differently sized datasets.

While its use in the neurosurgical field is still dawning, ML can potentially assist physicians in the pre-, intra-, and postsurgical settings. In the present study, the authors investigate whether an RF prediction model can accurately and reliably identify patients at high risk for IOL, testing the hypothesis that such a model would outperform multivariable logistic regression analysis.

Furthermore, to improve understanding, trust, and verification of the RF model predictions, a locally interpretable model-agnostic explanations (LIME) algorithm is employed.<sup>15</sup>

## Materials and Methods

### Data and Study Population

We collected and examined the data of 255 consecutive patients who underwent endoscopic transsphenoidal surgery for pituitary adenoma between January 2017 and February 2020 at the Department of Neurosurgery in Fondazione Policlinico Agostino Gemelli. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional

and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

For the purpose of the present study is to predict the IOL risk during transsphenoidal adenoma resection based on preoperative clinical and radiological data, 45 otherwise eligible patients were excluded because radiological images were not available. A total of 210 patients met our inclusion criteria and were included in our analysis.

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement guidelines were used to minimize the risk of bias during the development phase and to correctly validate the predictive ability of our ML models during the testing phase.<sup>16</sup>

A traditional statistical model (multivariable logistic regression) and a novel ML ensemble algorithm (the RF classifier) were trained to predict IOL: the two predictive models were then compared.

The design of the ML model herein presented is outlined in the flow chart (► Fig. 1). A stepwise description is provided hereinafter.

### Data Extraction

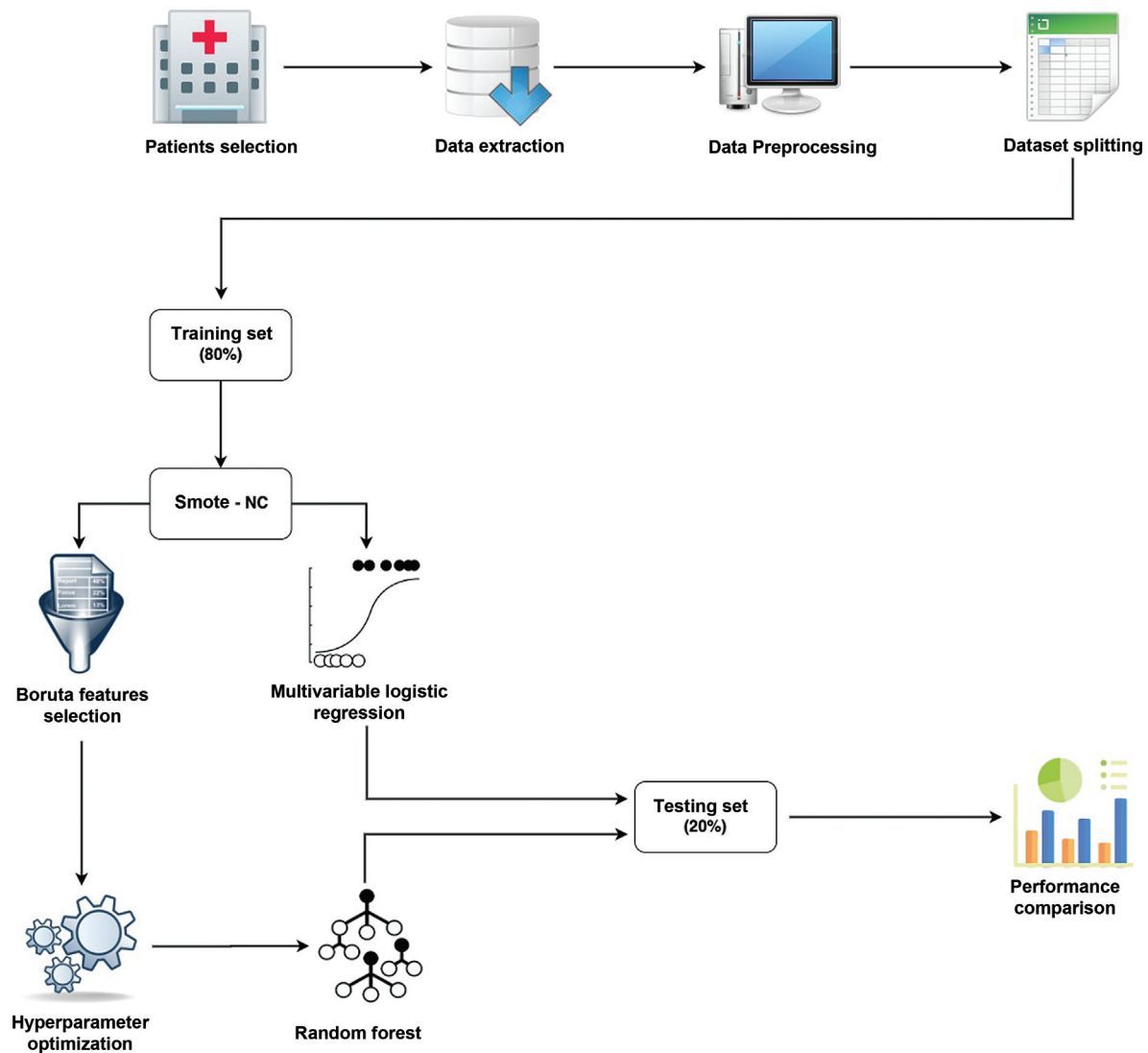
Input features included demographic, clinical, radiological, and surgical data from the database of 210 patients. The following patient's variables were retrieved: sex, age (years), adenoma secreting status (nonsecreting, adrenocorticotropic hormone [ACTH], growth hormone [GH], prolactin [PRL], and thyroid-stimulating hormone [TSH]), previous surgery (yes vs. no), preoperative pharmacological therapy (yes vs. no), maximum tumor diameter in mm (X, Y, and Z axes for laterolateral, craniocaudal, and anteroposterior, respectively), intercarotid distance (measured on T1-weighted [W] gadolinium-enhanced magnetic resonance imaging (MRI) at the level of the horizontal C4 segment of the internal carotid artery), Hardy's grade (sellar and suprasellar), Knosp's grade, and osteodural invasiveness (evaluated on a coronal T2-weighted image by a board-certified neuroradiologist with >10 years of experience).

### Features Selection and Training of The Random Forest Classifier

Choosing the most appropriate ML model strongly depends on the number of patients and the type of variables in the dataset, as well as the type of outcome of interest (binary vs. continuous).

Because of the nature of the variables included in our dataset, the number of patients and the outcomes of interest, a RF classifier was deemed the most appropriate.<sup>17–19</sup> Features selection for the RF classifier was performed using Boruta (v.0.3).<sup>20,21</sup> Boruta's initializing parameters are reported in **Supplementary Material S1** (available in the online version).

Issues deriving from the imbalanced nature of our dataset were explored. As widely acknowledged, when training with imbalanced data, ML algorithms tend to learn preferentially from the majority outcome class than the minority outcome class, these results in predictive models with limited



**Fig. 1** Flow chart describing the workflow behind the random forest model development and evaluation.

generalizability. To solve this issue and in line with previous experience,<sup>22</sup> a new balanced dataset was created applying the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC) to the original training dataset.<sup>23</sup>

The cohort of patients was randomly split into training and hold-out test set by an 80:20 ratio. The two sets were cross-checked for comparable class distribution. A grid search with five-fold cross-validation was used on the training set for hyperparameters tuning of the RF model. The best performing hyperparameters specific for the RF classifier (e.g.:  $n$  of estimators, learning rate, max depth, and others) are reported in **Supplementary Material S1** (available in the online version).

RF classifier was finally trained with the optimized hyperparameters setting on the balanced training set.

#### Internal Validation of the Random Forest Classifier

Once trained on the training set, the RF classifier was subsequently evaluated on the hold-out test set which was not employed in any form for hyperparameter optimization.

The RF model proceeds by identifying those patients who, based on their preoperative characteristics, are more prone to develop intraoperative CSF leakage during the transphenoidal resection of the pituitary adenoma. The ML model output estimates the probability of intraoperative CSF leakage occurrence ranging from 0 to 100%. IOL was predicted in patients with an estimated IOL probability  $>50\%$ .

#### Performance Metrics Evaluation and Comparison with Conventional Statistics

To validate the method, the RF classifier's performances were compared with the one achieved by a "classical" multivariable logistic regression model. The association between patients' variables and outcome of interest was explored; for categorical variables, with the Chi-square test, using the Fisher's exact test when appropriate; for continuous variables with the Mann-Whitney  $U$ -test. A  $p$ -value cut-off of 0.05 with Holm-Bonferroni correction was applied, thus shielding against type-1 error in the setting of multiple comparisons (**Table 1**). All covariates that were significantly

**Table 1** Baseline parameters of the 210 included patients

Parameter	Overall (n = 210)	No CSF leak (n = 165)	CSF leak (n = 45)	Uncorrected p-value	Holm–Bonferroni corrected p-value
Sex (female)	104 (49.5%)	83 (50.3%)	21 (46.7%)	0.665	>0.999
Age (y)	53.0 (15.1)	51.6 (15.3)	58.2 (13.4)	0.005	0.035 <sup>a</sup>
Secreting status	–	–	–	0.002	0.016 <sup>a</sup>
ACTH	23 (11%)	21 (12.7%)	2 (4.4%)	0.137	0.274
GH	34 (16.2%)	33 (20%)	1 (2.2%)	0.009	0.040 <sup>a</sup>
Nonfunctioning	125 (59.5%)	88 (53.3%)	37 (82.2%)	0.026	0.049 <sup>a</sup>
PRL	24 (11.4%)	21 (12.7%)	3 (6.7%)	0.286	0.572
TSH	4 (1.9%)	2 (1.2%)	2 (4.4%)	0.164	0.328
Previous surgery (yes vs. no)	33 (15.7%)	27 (16.4%)	6 (13.3%)	0.621	>0.999
Preoperative pharmacotherapy (yes vs. no)	26 (12.4%)	23 (13.9%)	3 (6.7%)	0.189	0.378
X	19.0 (8.9)	18.3 (9.3)	21.7 (6.2)	0.001	0.011 <sup>a</sup>
Y	18.5 (10.3)	17.2 (10.2)	23.3 (9.3)	<0.001	<0.001 <sup>a</sup>
Z	16.5 (7.2)	15.8 (7.2)	19.0 (6.7)	0.001	0.011 <sup>a</sup>
Intercarotid distance	20.2 (4.7)	20.6 (4.8)	18.8 (3.9)	0.009	<0.001 <sup>a</sup>
Knosp's grade	1.8 (1.2)	1.6 (1.2)	2.2 (0.9)	0.001	0.011 <sup>a</sup>
Hardy's (sellar) grade	1.9 (1.1)	1.7 (1.1)	2.5 (0.8)	<0.001	<0.001 <sup>a</sup>
Hardy's (suprasellar) grade	1.3 (1.3)	1.1 (1.3)	2.0 (0.9)	<0.001	0.001 <sup>a</sup>
Osteodural invasiveness	67 (31.9%)	46 (27.9%)	21 (46.7%)	0.017	0.045 <sup>a</sup>

Abbreviations: ACTH, adrenocorticotropic hormone; CSF, cerebrospinal fluid; GH, growth hormone; PRL, prolactin; TSH, thyroid-stimulating hormone.

Note: Characteristics of patients who experienced intraoperative CSF leaks and those who did not were compared at univariate analysis. Categorical and continuous variables are respectively reported as number of patients (%) and mean ( $\pm$ standard deviation).

<sup>a</sup>Significant at  $p \leq 0.05$  after Holm–Bonferroni correction;

associated with IOL in the inferential analysis were included in the predictive analysis of the multivariable logistic regression model (**►Table 2**).

A limit of this widely adopted statistical methodology is that it considers only the univariate association and correlation between the independent variables (e.g., diameter, sex, age, Knosp's grade, and others) and the outcome of interest while missing the valuable predictive value provided by the linear combination of patient's variables.

The RF classifier and the multivariable logistic regression model predictive capabilities were extensively evaluated and compared taking into consideration the following performance metrics:

- Area under the receiving operative characteristics (AUC-ROC).
- Accuracy.
- Sensitivity or recall.
- Positive predictive value (PPV).
- Negative predictive value (NPV) or precision.
- False-positive rate (FPR).
- F-1 score.

Mean value and 95% bootstrap CI were computed for each of the above-mentioned performance metrics across training and testing set by repeated cross-validation.<sup>24</sup>

### Software

All the statistical analyses were performed in Jupyter Notebook, using Python v.3.7.6 (<https://www.python.org/>). The Python packages used for this study included: “Scikit-learn” to develop and train the RF models and the multivariable logistic regression model; “Numpy” for Excel dataset handling; “imbalanced-learn” to solve class imbalances problem; “Sci-py” to perform univariable statistical association test; “Statsmodels” to perform multivariate analysis; “Boruta” to perform recursive features selection; “LIME” v.0.2.0.1 to interpret the ML model, “Bootstrap” v.4.5.3 to prototype the web application user interface.

### Results

A total of 210 consecutive patients operated between January 2017 and February 2020 were included. Their baseline clinical characteristics are reported in **►Table 1**.

**Table 2** Multivariate logistic regression analysis

Parameter	Odds ratio	95% CI	p-Value
Age	0.999	0.963–1.035	0.944
Secreting status	–	–	–
ACTH	1	–	–
GH	0.173	0.011–2.722	0.212
PRL	3.453	0.518–23.021	0.200
TSH	0.887	0.099–7.970	0.915
Nonfunctioning	11.268	0.617–205.798	0.102
X	1.011	0.910–1.122	0.843
Y	1.080	1.000–1.166	0.048
Z	0.924	0.815–1.046	0.212
Inter-carotid distance	0.725	0.632–0.830	<0.001
Knosp's grade	1.308	0.731–2.340	0.365
Hardy's (sellar) grade	1.175	0.609–2.267	0.631
Hardy's (suprasellar) grade	1.590	1.000–2.528	0.050
Osteodural invasiveness	2.273	0.940–5.494	0.068

Abbreviations: ACTH, adrenocorticotrophic hormone; CI, confidence interval; GH, growth hormone; PRL, prolactin; TSH: thyroid-stimulating hormone.

Note: All parameters showing a statistically significant association ( $p \leq 0.05$ ) at univariate analysis were included in the multivariate logistic regression analysis.

Intraoperative CSF leaks occurred in 45 patients (21.4%) of which 34 were classified as Park's grade 1 and 11 were classified as Park's grade 2.<sup>25</sup> Four patients with IOL (8.9%, three Park's grade 1 and one Park's grade 2), further developed a CSF leakage in the postoperative setting.

Out of 165 patients who did not develop IOL, 2 (1.2%) developed a CSF leakage in the postoperative setting.

In patients who did not develop IOL, sellar floor repair was performed by placing a sheet of Spongostan (Ethicon Inc., Johnson & Johnson Medical NV, Belgium) and adding fibrin glue and autologous bone/cartilage; depending on individual anatomy, predicted extent of sellar opening and surgeon's

preference, Hadad–Bassagasteguy flap was harvested in the initial stages of surgery and used to seal the surgical site.<sup>26</sup>

In patients who developed IOL, additional steps were taken to ensure watertight closure, depending on the Park's grade and on surgeon's preference; these variably included Hadad–Bassagasteguy flap, abdominal fat grafting, and gasket seal reconstruction, combined with the standard reconstruction technique.<sup>27</sup>

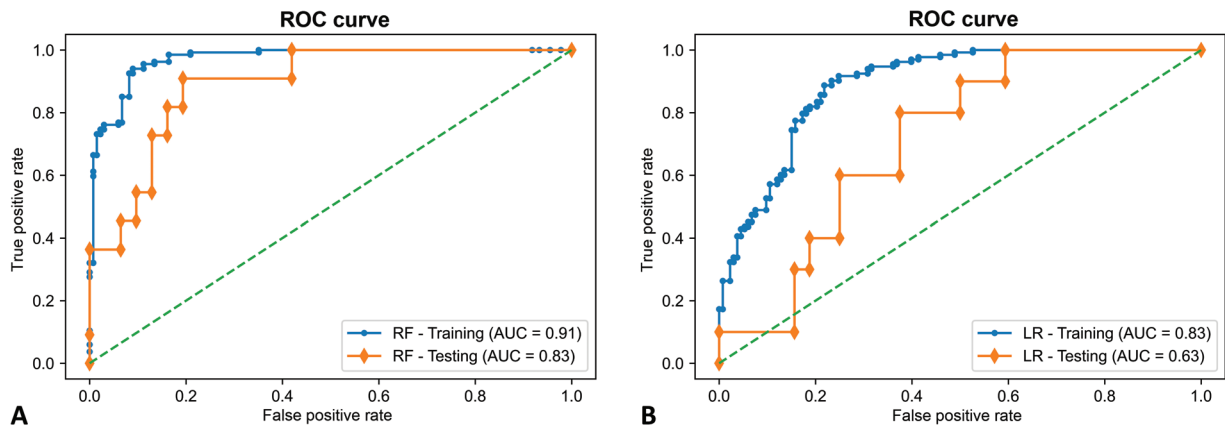
Of the six total patients who developed a CSF leakage in the postoperative setting, four were successfully treated with lumbar drain insertion, while two patients required an endoscopic procedure with autologous fat graft, autologous fascia lata graft and Hadad–Bassagasteguy flap.

**Table 3** Performance metrics on the training and testing set for both random forest classifier and multivariable logistic regression are reported

Performance metrics	Random Forest Classifier		Multivariable logistic regression		Improvement (%)
	Training	Testing	Training	Testing	
AUC	0.91 (0.89–0.94)	0.83 (0.78–0.86)	0.83 (0.79–0.84)	0.63 (0.57–0.65)	32
Accuracy	91% (86%–94%)	83% (80%–85%)	84% (79%–86%)	64% (61%–66%)	30
Sensitivity	98% (94%–99%)	82% (79%–83%)	86% (83%–87%)	60% (55%–62%)	37
Specificity	84% (82%–89%)	84% (81%–86%)	80% (75%–82%)	66% (62%–69%)	27
PPV (precision)	86% (82%–89%)	64% (63%–68%)	82% (79%–84%)	35% (33%–38%)	83
NPV	97% (94%–99%)	93% (89%–95%)	86% (85%–88%)	84% (79%–86%)	11
False positive rate	14% (12%–17%)	16% (13%–18%)	20% (18%–25%)	34% (30%–36%)	–53
F1 score	0.91 (0.88–0.93)	0.72 (0.70–0.75)	0.84 (0.81–0.86)	0.44 (0.40–0.49)	64

Abbreviations: AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

Note: Metrics improvement on the testing set is listed; 95% bootstrap confidence interval is reported in brackets.



**Fig. 2** Mean Receiver Operating Characteristic (ROC) curves on both training and testing sets using the Random Forest model (A) and multivariable logistic regression (B) to predict intraoperative CSF leakage in patients with pituitary adenoma. AUC, area under the curve; LR, logistic regression; RF, Random Forest; ROC, receiver operating characteristics.

**Random Forest Classifier Performances**

Features selected for the RF classifier via Boruta were age, Knosp’s grade, Hardy’s grade (both sellar and suprasellar), tumor diameter (on X, Y, and Z axes), intercarotid distance, and secreting status (nonfunctioning and GH secreting).

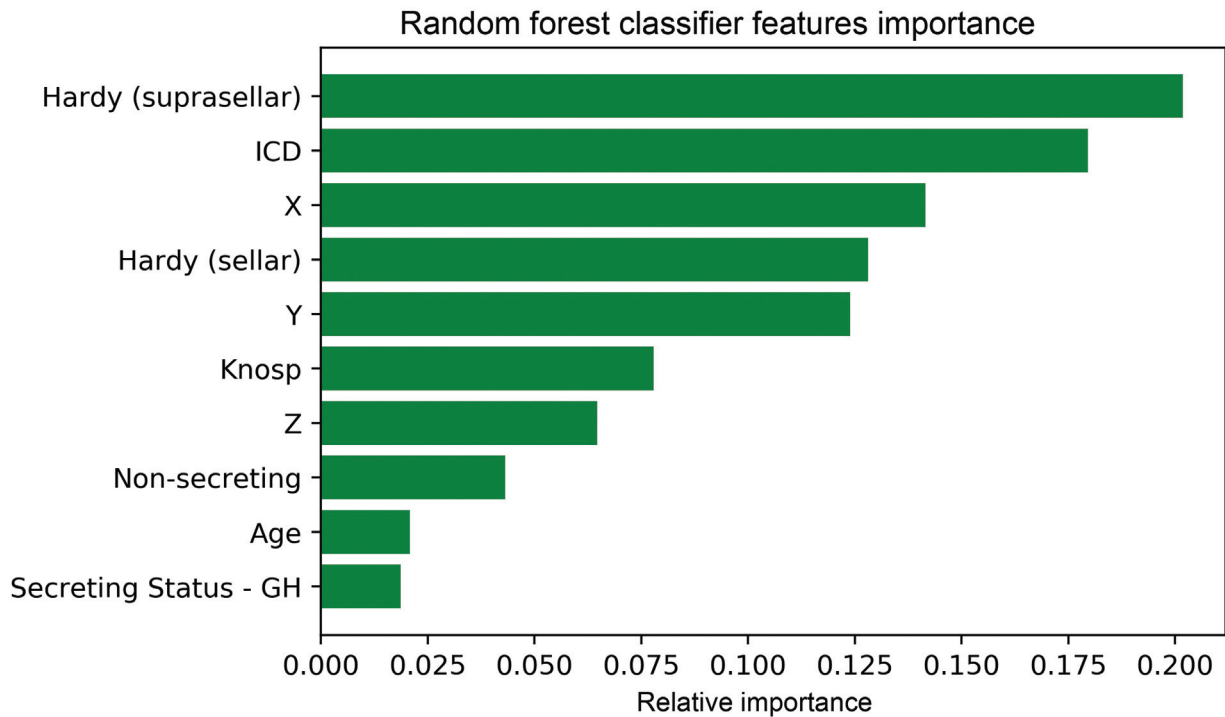
The results of the RF prediction model are reported in ►Table 3. Comparing the performance achieved on the training set with those of the testing set, the RF model demonstrated internal validity and minimal overfitting.

This model accurately classified 83% of patients in the hold-out test set and achieved an AUC of 0.83 (95% CI: 0.78; 0.86) demonstrating adequate discriminative ability

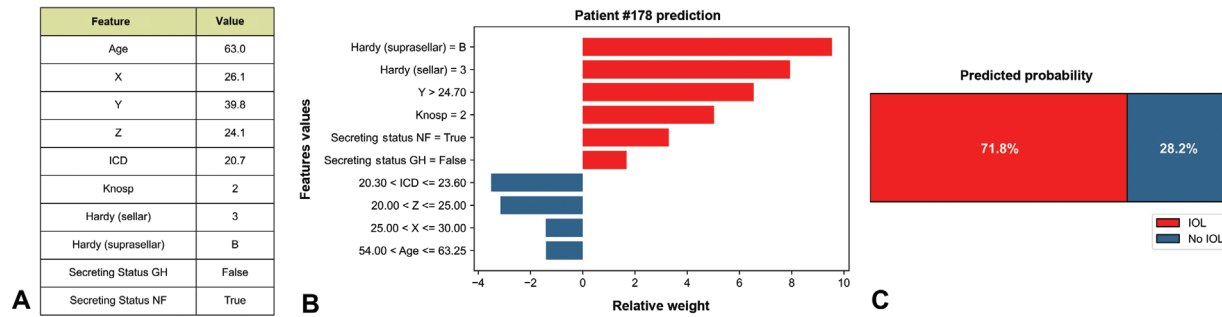
(►Fig. 2). Notably, the RF model achieved an NPV rate of 93%, indicating high reliability in correctly identifying patients at minimal risk for IOL. Evaluation of the model on the test set corresponds to internal validation, providing reliable expectation on the model’s performance on new, external data; if similar results were achieved by external validation, it would allow the introduction of the model in clinical practice.

**Model Interpretation**

Relative features importance plot for the RF model is reported in ►Fig. 3. The parameters with the strongest



**Fig. 3** Permutation features importance plot providing a visual representation of the relative predictive contribution of each selected variable to minimize the prediction error of the random forest model. The relationship between permutation feature importance and outcome of interest is nonlinear and cannot be interpreted directionally with respect to their influence on outcomes, nor can they be used to generate specific cutoff or threshold values. GH, growth hormone; ICD, Intercarotid distance; NF, nonfunctioning.



**Fig. 4** Results of local interpretable model-agnostic explanation (LIME) with Random Forest classifiers applied to one correctly predicted patient with intraoperative cerebrospinal fluid (CSF) leakage (IOL). The figure reveals the role of various features in shaping the risk of IOL in each patient. (A) Patient's characteristics. (B) Features contributions on predicted probabilities (red = risk factor; blue protective factor). (C) Predicted probability of IOL. GH, growth hormone; ICD, Intercarotid distance.

predictive value were suprasellar Hardy's grade, Inter-carotid distance (ICD), and tumor diameters (X, Y, and Z axes).

We further introduced LIME to quantify the features contribution and polarity for each patient, thus providing an interpretable relationship between patient's characteristics and RF model prediction. An example is illustrated in ►Fig. 4.<sup>15</sup> Understanding the reason behind both correct and incorrect model predictions can increase clinicians trust in model behavior and performance.

### Logistic Regression Performance

Out of the 13 available features, 10 reported a  $p < 0.05$  at univariate analysis and were included as input variables in the logistic regression model (►Table 2). Of these, ICD and tumor diameter (Y) resulted independently associated with IOL, while suprasellar Hardy's grade and osteodural invasiveness trended toward significance. On the hold-out test set, the multivariable logistic regression model achieved, AUC of 0.63 (95% CI: 0.57; 0.65; ►Fig. 2), indicating poor consistency and scarce reliability (►Table 3).

### Performance Metrics Evaluation

By leveraging the predictive value resulting from the combination of independent variables, RF classifier outperformed multivariable logistic regression in successfully identifying patients at high risk for intraoperative leaks (►Table 3). Improvement was recorded across all evaluation metrics. Notably, the positive predictive value increased by +83%, the NPV increased by +11%, and the false-positive rate dropped by -53%.

### Discussion

Pituitary adenomas represent approximately 16% of all newly diagnosed brain tumors and are among the most common primary central nervous system tumors in the United States.<sup>28</sup> Moreover, they are the second most common nonmalignant brain tumor with surgical resection as a potentially curative treatment.

TSS is currently the gold standard for the treatment of pathologies of the sellar region, with significant improvement in long-term clinical outcomes and a marked reduction in the duration of hospitalizations, compared with the

traditional microsurgical technique.<sup>29</sup> Noticeably, among the main TSS-related risks, CSF leaks represent one of the most common complications that the pituitary surgeon has to face.<sup>5</sup>

IOL represents a rather common situation during TSS, occurring in up to 37.4% of TSS interventions, as reported by Strickland et al<sup>5</sup>; it can be expected in case of lesions with evident intracranial extension, or it can occur during surgery due to the presence of tumor adhesions or local invasion or involuntary laceration of the sellar diaphragm and the arachnoid plane during the surgical manipulation. Several studies have evaluated the risk factors responsible for an increased incidence of IOL, without reaching univocal conclusions; some of the reported associated variables include larger tumor size, nonsecreting status, previous surgery, suprasellar extension, and higher body mass index (BMI).<sup>5,11,30,31</sup>

Interestingly, in our study the variables selected by Boruta overlapped with the statistically significant variables ( $p < 0.05$ , as identified by multivariable logistic regression) except for osteodural invasiveness. This finding, coupled with the relative importance attributed to each variable (►Fig. 2), highlights that statistical significance alone is of limited value, only by considering the interactions between the several variables a model can achieve a high AUC and rate the contribution of each variable to the prediction. For instance, age seems to be tightly linked to the development of IOL ( $p = 0.009$ ); its contribution, however, is defined as marginal in our RF-based prediction model. This finding contrasts with what found by Staartjes et al whose neural network-based prediction model defines age as one of the most significant variables together with suprasellar Hardy's grade and previous surgery.<sup>14</sup> As literature on this association is scarce, further studies are required to define whether a link exists between age and IOL development.

In our study, the suprasellar Hardy grade contributes the most to the development of IOL, similarly to what previously described.<sup>14,30</sup> Similarly, tumor height was found to be more predictive of intraoperative CSF leak than anteroposterior and laterolateral dimensions, emphasizing the importance of considering tumor dimensions independently. Tumors with greater craniocaudal extension develop incompetence of the sellar diaphragm secondary to sellar expansion, leading to

exposed arachnoid that is at risk for thinning or developing defects, thereby increasing the risk of CSF leak.<sup>8,32</sup>

A larger ICD was found to act as a protective factor in the multivariable logistic regression (► **Table 2**), and it was selected by Boruta as the second most relevant feature in determining IOL development. While this association has not been previously reported, a greater ICD may lower the risk of diaphragm violation during surgery by virtue of the associated greater diaphragm diameter and surface.

The association between IOL and the nonsecreting status of the adenomas can be explained by considering the natural history of these tumors which tend to be larger at surgery because they are diagnosed after the onset of mass effect symptoms.<sup>33</sup> Tumor dimensions have been previously reported as reliable predictors of IOL.<sup>5,11,30</sup> While cavernous sinus invasion, as defined by the Knosp grade, plays a significant role in our prediction model, it has been previously described as a marginal determinant of IOL by Staartjes et al and as completely unrelated to IOL by Patel et al.<sup>14,30</sup> No association between prior surgery and the outcome was identified by means of either standard statistics or ML-based analysis; the literature in this regard is discordant, as Przybylowski et al reported no difference in IOL rates between primary and revision procedures, while the prediction model by Staartjes et al selected it as the third most relevant feature, probably as a result of fibrotic scars, and difficult recognition/dissection of the sellar structures.<sup>4,14</sup>

Among radiomic features that may predispose to IOL development, tumor texture stands out; the signal intensity of T2-W imaging and apparent diffusion coefficient images can have prediction value for texture of pituitary adenomas. Soft tumors are easily removed by suctioning, fibrous tumors are more difficult to excise and often require a second-stage operation, stereotactic radiosurgery, or transcranial approaches; it is only reasonable that harder consistency tumors would favor the development of IOL. Nevertheless, the accuracy of tumor texture prediction based on MRI signal is 70%,<sup>34,35</sup> for which reason we refrained from using this variable in our algorithm. Should a greater accuracy of radiomic features, tumor texture prediction included, be achieved in future, these could be implemented in existing prediction models.

ML has already been used in the literature to predict the risk of IOL and of the likelihood of GTR during and after transphenoidal surgery.<sup>13,14</sup> However, while those articles deployed artificial neural networks (ANNs) which are computational models based on the functioning of biological neural networks that can be used to model nonlinear statistical data and to reveal patterns,<sup>36,37</sup> we used an RF algorithm.

Since no single algorithm works best across all possible scenarios, the performance of ML algorithms varies widely depending on the application and the dimensionality of the dataset. Accordingly, the weakness of an approach can lead to avoid a specific algorithm in a specific context. In these cases, choosing an algorithm before starting the project is warranted.

Both ANNs and RFs have the ability to model linear, as well as complex nonlinear relationships. ANNs can lead to signif-

icant advantages in the analysis of complex data, such as image classification, speech recognition, and others; however, there's evidence, both in neurosurgery and in other specialties, that RFs could outperform other predictive algorithms, ANNs included, in the analysis of tabular data.<sup>38-40</sup>

RFs include several advantages in the analysis of tabular data, they can be trained with a relatively small amount of data, while ANNs require more data to reach the same level of accuracy and they require less input preparation, as no feature normalization is required.<sup>41</sup> Finally, differently from ANNs, RFs can accurately predict the outcome even when part of the input values are missing<sup>41</sup>; this occurrence is extremely common in the analysis of tabular data, often leading to the exclusion of patients from analysis if standard statistical methods (e.g., multivariable logistic regression) or other ML algorithms like ANNs are used. RFs-based analysis can possibly overcome this limit and be the basis for future prediction models starting from tabular data.

Despite widespread adoption, ML models are often viewed as black boxes; in the absence of a transparent interpretation of the learning process or the outputs, the doctor is blind to the relationship between the clinical features and the predicted outcomes. Understanding the reason behind each prediction is crucial to build clinicians' trust in ML models, and to provide expert knowledge-based validation for the interpretation of ML model outputs.

The interpretability of an ML algorithm is generally defined as the ability of a human to understand the link between the features extracted by an artificial intelligence program and its predictions.<sup>42</sup>

LIME algorithm was introduced to provide an explanation on a case-by-case basis for the RF classifier prediction (► **Fig. 4**). LIME is an algorithm that can explain the predictions of any classifier or regressor in a faithful way by approximating it locally with an interpretable model.<sup>15</sup> An in-depth knowledge of what drives ML model prediction is necessary for an effective human-ML systems interaction.

Our study is the first to deploy an RF-based algorithm to predict IOL, and, with 210 recruited patients, it is one of the largest studies in the field of ML algorithms applied to pituitary surgery. Our ML-based prediction model outperformed multivariable logistic regression: by achieving an AUC of 0.83 (95% CI: 0.78; 0.86), it demonstrates a high discriminative ability and generalizability.

A web application user interface has been designed for the clinical deployment of our random forest model (► **Fig. 5**). For safety reasons, a publicly accessible version will be released only on successful fulfilment of the currently ongoing multicentric and prospective data collection and validation of the current model. External validation of our model is necessary to adjust for variability in surgical technique, as interindividual differences in surgical technique surely can impact the outcomes due to the long learning curve of transnasal endoscopy; all patients in our study were operated on by the same surgical team lead by one senior surgeon with more than 15 years of experience in endonasal endoscopic surgery; hence, we can reliably state that the same surgical technique was deployed in all cases.



**Figure 5 Data Summary:**

Age	Secreting status	Hardy sellar	Hardy suprasellar	Knosp	X (mm)	Y (mm)	Z (mm)	ICD (mm)	Estimated Probability	Risk Classification
41	GH	2	B	0	16	17	13	18	20%	Low risk
55	PRL	2	C	2	24	22	18	20	68%	Intermediate risk
68	Non-functioning	3	C	4	37	49	35	24	87%	High risk

**Fig. 5** Webapp prototype of the graphic user interface for the clinical deployment of the random forest prediction model. According to the estimated probability, each patient could be classified as being at low, intermediate or high risk of developing intraoperative CSF leaks following three risk ranges: 0–33, 34–66, and 67–100%. The risk prediction of three patients, randomly sampled from the current database, is represented. CSF, cerebrospinal fluid; GH, growth hormone; PRL, prolactin.

If extended to everyday clinical practice, this ML-based decision support tool may guide the surgeon in decision-making and surgical planning by identifying patients at risk for IOL, leading to reduced surgical time and lower costs; most importantly, it could achieve lower morbidity and risks for the patients in terms of surgery and anesthesia-associated complications and postoperative infections. While our model could provide an objective risk quantification, the role and intuition of the surgeon remain crucial to provide adequate patient care.

## Limitations

The major drawback of the study resides in the retrospective acquisition of data from a single tertiary care center; such feature potentially challenges the generalizability of the current version of our model in external patient populations. Though we used a hold-out validation technique to demonstrate the generalizability of the ensembles to data never used in training, demonstration of generalizability to a separate database, or to prospectively collected data, it would serve as a stronger validation. The variables included in our prediction model can be easily retrieved for each patient in most neurosurgical centers, such as demographics, diameters on the three axes, ICD, Knosp and Hardy grades as measured on T1-W gadolinium-enhanced or T2-W sequences, where appropriate; however, poor interrater reliability (as in the case of Knosp's and Hardy's grades) may lead to poorer prediction performance. Furthermore, there are factors that may concur in the development of IOL that can't be included in a prediction model, such as individual surgical techniques, which may differ across different neurosurgical centers or even different neurosurgical teams based in the same center. For these reasons, external validation of the model is required to confirm its predictive capacity, possibly

heralding the implementation of a free web-based version of the model in clinical practice. As ML models evolve continuously with the use and accrual of new data, it can be predicted that the diffusion of this model in several neurosurgical centers may allow, in the future, the creation of multiple center-specific versions which adjust the results for individual surgical style and personal Knosp's and Hardy's grading.

## Conclusion

An RF-based prediction model was trained and internally validated to identify patients at risk for intraoperative CSF leakage; the AUC was 0.83 (95% CI: 0.79; 0.84) and the NPV value was 93%. The prediction model achieved superior results in comparison with conventional statistical methods whose AUC was 0.63 (95% CI: 0.57; 0.65); this finding supports the role of ML algorithms as auxiliary tool to aid physicians in clinical practice, hopefully resulting in reduced health care costs and improved patient care. While the results of our study seem encouraging, our prediction model needs to successfully fulfil the currently ongoing multicentric and prospective external validation before being safely introduced in everyday clinical practice.

## Abbreviations

AUC-ROC	area under the curve receiver operating characteristics
CSF	cerebrospinal fluid
FPR	false positive rate
RF	random forest
IOL	intraoperative CSF leakage
ML	machine learning

LIME	locally interpretable model-agnostic explanations
NPV	negative predictive value
PPV	positive predictive value
SMOTE-NC	synthetic minority over-sampling technique for nominal and continuous
TPR	true positive rate
TRIPOD	transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
TSS	transsphenoidal surgery

#### Ethical Approval

Ethical approval was waived by the local Ethics Committee in view of the retrospective nature of the study and all the procedures being performed were part of the routine care.

#### Consent to Participate

Informed consent was obtained from all individual participants included in the study.

#### Availability of Data and Material

The dataset that supports the findings of this study are available from the corresponding author, M.G., on request.

#### Code Availability

The source code employed to develop the herein presented machine learning model is available at the following GitHub repository: <https://github.com/valerio-mc/ML-fistola-pituitary>.

#### Conflict of Interest

None declared.

## References

- Chen CJ, Ironside N, Pomeranec JJ, et al. Microsurgical versus endoscopic transsphenoidal resection for acromegaly: a systematic review of outcomes and complications. *Acta Neurochir (Wien)* 2017;159(11):2193–2207
- Dhandapani S, Singh H, Negm HM, Cohen S, Anand VK, Schwartz TH. Cavernous sinus invasion in pituitary adenomas: systematic review and pooled data meta-analysis of radiologic criteria and comparison of endoscopic and microscopic surgery. *World Neurosurg* 2016;96:36–46
- Fatemi N, Dusick JR, Mattozo C, et al. Pituitary hormonal loss and recovery after transsphenoidal adenoma removal. *Neurosurgery* 2008;63(04):709–718, discussion 718–719
- Przybylowski CJ, Dallapiazza RF, Williams BJ, et al. Primary versus revision transsphenoidal resection for nonfunctioning pituitary macroadenomas: matched cohort study. *J Neurosurg* 2017;126(03):889–896
- Strickland BA, Lucas J, Harris B, et al. Identification and repair of intraoperative cerebrospinal fluid leaks in endonasal transsphenoidal pituitary surgery: surgical experience in a series of 1002 patients. *J Neurosurg* 2018;129(02):425–429
- Zhang C, Ding X, Lu Y, Hu L, Hu G. Cerebrospinal fluid rhinorrhoea following transsphenoidal surgery for pituitary adenoma: experience in a Chinese centre. *Acta Otorhinolaryngol Ital* 2017;37(04):303–307
- Karnezis TT, Baker AB, Soler ZM, et al. Factors impacting cerebrospinal fluid leak rates in endoscopic sellar surgery. *Int Forum Allergy Rhinol* 2016;6(11):1117–1125
- Mehta GU, Oldfield EH. Prevention of intraoperative cerebrospinal fluid leaks by lumbar cerebrospinal fluid drainage during surgery for pituitary macroadenomas. *J Neurosurg* 2012;116(06):1299–1303
- Conger A, Zhao F, Wang X, et al. Evolution of the graded repair of CSF leaks and skull base defects in endonasal endoscopic tumor surgery: trends in repair failure and meningitis rates in 509 patients. *J Neurosurg* 2018;130(03):861–875
- Zhou Q, Yang Z, Wang X, et al. Risk factors and management of intraoperative cerebrospinal fluid leaks in endoscopic treatment of pituitary adenoma: analysis of 492 patients. *World Neurosurg* 2017;101:390–395
- Jakimovski D, Bonci G, Attia M, et al. Incidence and significance of intraoperative cerebrospinal fluid leak in endoscopic pituitary surgery using intrathecal fluorescein. *World Neurosurg* 2014;82(3,4):e513–e523
- Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg* 2018;109:476–486.e1
- Staatjes VE, Serra C, Muscas G, et al. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. *Neurosurg Focus* 2018;45(05):E12
- Staatjes VE, Zattra CM, Akeret K, et al. Neural network-based identification of patients at high risk for intraoperative cerebrospinal fluid leaks in endoscopic pituitary surgery. *J Neurosurg* 2019 (e-pub ahead of print). Doi: 10.3171/2019.4.JNS19477
- Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You?": Explaining the predictions of any classifier. Accessed December 2, 2021 at: <https://aclanthology.org/N16-3020.pdf>
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(01):55–63. Doi: 10.7326/M14-0697
- Breiman L. Random forests. *Mach Learn* 2001;45:5–32
- Liaw A, Wiener M. Classification and regression by random forest. *R News* 2002;2:18–22
- Cutler A, Cutler DR, Stevens JR. Random forests. In: Zhang C, Ma Y, eds. *Ensemble Machine Learning*. Boston, MA: Springer; 2012: 157–175
- Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010;36(11):1–13
- Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform* 2019;20(02):492–503
- Sakr S, Elshawi R, Ahmed AM, et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercise testing (FIT) project. *BMC Med Inform Decis Mak* 2017;17(01):174
- Chawla NV, Bowyer K, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16(01):321–357
- DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Institute of Mathematical Statistics* 1996;11(03):189–228
- Park JH, Choi JH, Kim YI, Kim SW, Hong YK. Modified graded repair of cerebrospinal fluid leaks in endoscopic endonasal transsphenoidal surgery. *J Korean Neurosurg Soc* 2015;58(01):36–42
- Hadad G, Bassagasteguy L, Carrau RL, et al. A novel reconstructive technique after endoscopic expanded endonasal approaches: vascular pedicle nasoseptal flap. *Laryngoscope* 2006;116(10):1882–1886
- Cavallo LM, Solari D, Somma T, Cappabianca P. The 3F (fat, flap, and flash) technique for skull base reconstruction after endoscopic endonasal suprasellar approach. *World Neurosurg* 2019; 126:439–446

- 28 Ostrom QT, Gittleman H, Xu J, et al. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2009–2013. *Neuro Oncol* 2017;18 (Suppl 5):v1–v75
- 29 Komotar RJ, Starke RM, Raper DM, Anand VK, Schwartz TH. Endoscopic endonasal compared with microscopic transsphenoidal and open transcranial resection of giant pituitary adenomas. *Pituitary* 2012;15(02):150–159
- 30 Patel PN, Stafford AM, Patrinely JR, et al. Risk Factors for Intraoperative and postoperative cerebrospinal fluid leaks in endoscopic transsphenoidal sellar surgery. *Otolaryngol Head Neck Surg* 2018;158(05):952–960
- 31 Xue H, Wang X, Yang Z, Bi Z, Liu P. Risk factors and outcomes of cerebrospinal fluid leak related to endoscopic pituitary adenoma surgery. *Br J Neurosurg* 2020;34(04):447–452
- 32 Cavallo LM, Messina A, Cappabianca P, et al. Endoscopic endonasal surgery of the midline skull base: anatomical study and clinical considerations. *Neurosurg Focus* 2005;19(01):E2
- 33 Fleseriu M, Karavitaki N. Non-functioning pituitary adenomas, not all the same and certainly not boring!. *Pituitary* 2018; 21:109–110
- 34 Snow RB, Johnson CE, Morgello S, Lavyne MH, Patterson RH Jr. Is magnetic resonance imaging useful in guiding the operative approach to large pituitary tumors? *Neurosurgery* 1990;26(05):801–803
- 35 Wei L, Lin SA, Fan K, Xiao D, Hong J, Wang S. Relationship between pituitary adenoma texture and collagen content revealed by comparative study of MRI and pathology analysis. *Int J Clin Exp Med* 2015;8(08):12898–12905
- 36 Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. *J Neurol Neurosurg Psychiatry* 2015;86(03):251–256
- 37 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521 (7553):436–444
- 38 Fernandez-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* 2014;15:3133–3181
- 39 Nawar S, Mouazen AM. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors (Basel)* 2017;17(10):2428. Doi: 10.3390/s17102428
- 40 Senders JT, Staples P, Mehrtash A, et al. An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neurosurgery* 2020;86(02): E184–E192
- 41 Ahmad MW, Mourshed M, Rezguy Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build* 2017; 147:77–89
- 42 Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell* 2020;2(03):e190043