Mohammadzadeh *et al.*
BMC Endocrine Disorders (2025) 25:158
https://doi.org/10.1186/s12902-025-01955-8

BMC Endocrine Disorders

RESEARCH

Open Access

Prediction of recurrence after surgery for pituitary adenoma using machine learning- based models: systematic review and meta-analysis

Ibrahim Mohammadzadeh[1,12*†], Bardia Hajikarimloo[2†], Behnaz Niroomand[1], Nasira Faizi[3], Pooya Eini[4], Mohammad Amin Habibi[5], Alireza Mohseni[6], Mohammadmahdi Sabahi[7], Abdulrahman Albakr[7,8], Michael Karsy[9] and Hamid Borghei-Razavi[7,10,11*]

Abstract

Background Predicting pituitary adenoma (PA) recurrence after surgical resection is critical for guiding clinical decision-making, and machine learning (ML) based models show great promise in improving the accuracy of these predictions. These models can provide valuable insights to surgeons and oncologists, helping them tailor personalized treatment plans, enhance patient prognostication, and optimize follow-up strategies.

Methods We systematically searched PubMed, Scopus, Embase, Cochrane Library, and Web of Science databases until November 2024, applying PRISMA guidelines.

Results Out of 1240 studies screened, six met our eligibility criteria involving ML-based approaches to predict PA recurrence. The studies employed 12 different ML algorithms. Meta-analysis showed a pooled sensitivity of 0.87 [95% CI: 0.78–0.92], specificity of 0.86 [95% CI: 0.67–0.95], positive diagnostic likelihood ratio (DLR) of 6.32 [95% CI: 2.46–16.26], and negative DLR of 0.16 [95% CI: 0.1–0.25]. The diagnostic odds ratio (DOR) was 40.52 [95% CI: 13–126.27], and the diagnostic score was 3.7 [95% CI: 2.57–4.84]. The pooled AUC was 0.89 [95% CI: 0.86–0.92], indicating a high overall diagnostic performance. For the comparison between Logistic Regression (LR) and non-LR algorithms, LR-based algorithms exhibited numerically higher AUC and sensitivity; however, these differences were not statistically significant. Additionally, LR-based algorithms showed lower specificity, positive likelihood ratio, and diagnostic odds ratios, but the statistical tests did not provide strong evidence for meaningful differences.

Conclusion AI-based models show strong predictive power for recurrence in both functional and non-functional pituitary adenomas, with an average accuracy above 80%. However, the lack of external validation and the complexity of input data pose challenges, highlighting the need for rigorous validation with multi-center datasets and standardized imaging techniques to enhance clinical applicability.

†Ibrahim Mohammadzadeh and Bardia Hajikarimloo contributed equally to this work as the first author.

*Correspondence:
Ibrahim Mohammadzadeh
Ibrahim.mdz7777@gmail.com; ibrahim.mohammadzadeh@sbmu.ac.ir
Hamid Borghei-Razavi
borgheh2@ccf.org
Full list of author information is available at the end of the article

Keywords Pituitary adenoma, Recurrence, Machine learning, Artificial intelligence, Deep learning, Predictors

Introduction

Pituitary adenomas (PAs) constitute 10–25% of all intracranial tumors and are prevalent in approximately 17–20% of the general population [1, 2]. Based on their secretion status, PAs can be divided into functioning adenomas (FPA), which actively secrete hormones with or without mass effects, and nonfunctioning adenomas (NFPAs), which usually only have mass effects complications [3]. PAs can impose various morbidities on patients regarding hormonal dysfunction, including amenorrhea-galactorrhea in women and sexual dysfunction in men with prolactin (PRL)-secreting adenoma; acromegaly and gigantism in growth hormone (GH)-secreting adenomas; Cushing's disease (CD) in adrenocorticotropic hormone (ACTH)-secreting corticotrophinomas; and central hyperthyroidism in the rare TSH (thyrotropin)-secreting thyrotrophinomas [4, 5]. Nonfunctioning PAs may present with symptoms of mass effect, which include visual field defects, headaches, anterior hypopituitarism, and diabetes insipidus (DI) in infiltrative or suprasellar lesions [6]. Clinically nonfunctioning microadenomas (diameter < 10 mm) are unlikely to cause mass effects and are usually detected incidentally [6].

Surgical removal through microscopic or endoscopic transnasal transsphenoidal approaches is the primary treatment option; however, up to 40–50% of cases achieve complete resection in some series, and at least 10–20% of entirely resected tumors recur after 5–10 years [7]. Among patients with residual adenomas, 12–58% experience recurrence [8]. Recurrence of PAs significantly affects the quality of life of patients due to the impact of pituitary dysfunction, invasion-related risks, and increased risk with additional surgical or nonoperative treatment [8].

Artificial intelligence (AI) is an increasingly applicable field with many neuro-oncology applications [9]. AI models, including machine learning (ML), artificial neural networks (ANN), and deep learning (DL), have been developed to perform big data analysis and predictive analysis in medicine [10–14]. ML algorithms can be used to generate predictive models based on large datasets and iteratively learning. Studies show the efficient performance of ML-based models over some conventional models for a variety of radiological and clinical outcomes [15–17]. AI can potentially improve the diagnosis and treatment of brain tumors and provide a path toward personalized medicine and better patient outcomes [18–20]. We aimed to investigate the application of AI in forecasting recurrence in patients with previously treated PA. Through a systematic review, we evaluated the effectiveness and accuracy of AI algorithms in predicting the likelihood of tumor recurrence. The primary goal was to explore the accuracy of AI-based models as well as added value to clinical decision-making by providing reliable and precise predictions of recurrence.

Method

Study design

This systematic review and meta-analysis examined the current role of AI algorithms, including ML and DL, in forecasting recurrence following surgical resection PAs. The study follows the guidelines based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [21], and has been registered with PROSPERO ID: CRD42024627890.

Search strategy

Five major databases in medical research, including PubMed, Embase, Scopus, Cochrane Library, and Web of Science, were comprehensively searched using the following search terms: ("machine learning" OR "deep learning" OR "artificial intelligence") AND ("pituitary adenoma" OR "pituitary tumor" OR "ACTH-secreting" OR "GH-secreting" OR "TSH- secreting" OR "thyrotropin-secreting" OR "Cushing's disease" OR corticotrophinomas OR thyrotrophinomas OR adrenocorticotropic OR prolactinoma OR somatotropinoma OR "nonfunctional adenoma" OR "secretory adenoma") AND (recurrence OR residual OR regrowth). No restriction on publication type or language was applied and they were searched from their inception up to 27 November 2024. The first 100 results from Google Scholar were also reviewed as a supplementary search to ensure all relevant articles were included. Special syntax of each database such as the medical subject heading (Mesh) in PubMed and Emtree in Embase were used to retrieve the unique search strategy for each database. The complete search strategy syntax is provided in Supplementary file 1, Table S1.

Inclusion criteria

The eligible studies were selected according to the following criteria: 1) Study participants consisting of patients already diagnosed with PA and undergone endoscopic or microscopic TSS, 2) At least one year of follow-up after surgery, 3) Studies reporting the application of ML or DL algorithms in the prediction of the recurrence or regrowth of PA.

Exclusion criteria

Studies were excluded if they were non-English language, animal studies, commentaries, conference abstracts, case reports, or case-series with fewer than 15 patients, did not reporting sufficient data on our intended parameters (sensitivity, specificity, AUC, true positives (TP) and true negatives (TN)), and included patients younger than 18 years old.

Study selection

After a systematic literature search, the articles were exported from databases into EndNote software (version 21). After resolving the duplicates, the titles and abstracts of the retrieved records were reviewed and screened independently by two authors (B.N. and P.E.). A full-text assessment resolved conflicts. Subsequently, the full-text screening of the included or possible articles was performed by two authors (N.F. and B.H.). At this stage, the disagreements were resolved by a third reviewer (I.M.). No automatic tools were used during the selection process.

Data extraction

Two reviewers (B.N. and P.E.) separately collected the data from the included articles. A pre-designed Excel form was utilized to submit data from each study. From each article the following data were collected: study design, sample size, average age, gender composition, type of the tumor (adenoma), criteria for recurrence, time to recurrence, duration of follow-up, modality of imaging, input characteristics, validation method, selected features, number of final features, type of AI model(s), the best-performing model and its sensitivity, specificity, accuracy, precision, F1 score, and AUC.

Risk of bias and applicability assessment

The risk of bias and applicability of the terminally included articles was assessed using the PROBAST (Prediction Model Risk of Bias Assessment Tool) [22]. PROBAST was primarily intended for use in systematic reviews, but generally, it can be used more in critically evaluating prediction model studies. The applicability of the prediction models determines the compatibility of the included studies with the research question. This tool examines four key domains for assessing the risk of bias, including participants, predictors, outcome, and analysis, and the same except for the analysis domain for applicability. Each domain was ranked as low, high, or unclear risk.

Statistical analysis

The true positive, true negative, false positive, and false negative values were derived based on the sensitivity and specificity of each algorithm in every study. A diagnostic meta-analysis model was then applied using these values to synthesize data. Statistical analysis for all data was performed using the MIDAS module in STATA version 17. The primary objective of this review was to identify the top-performing ML algorithm from each study. Additionally, a secondary analysis was conducted on all reported algorithms to provide a more comprehensive overview.

Result

The initial search identified 1461 records, of which 207 were duplicates and non-English (Fig. 1). After resolving them, 1254 records were screened by title and abstract, and 11 studies were eligible for full-text assessment. Finally, six studies published between 2019 and 2024 were included in the systematic review, with their sample sizes ranging from 27 to 354, comprising a total of 1188 subjects.

Female gender accounted for 56.23% of the sample (female/male ratio: 1.28), and the mean age was 44.3 years (Table 1). Three studies included CD patients (n = 719), one of them also surveyed participants with acromegaly (n = 191), and three of studies examined NFPA (n = 267). furthermore, three were from China (50%), and one each from the US, Russia, and Brazil (13.33%) (Table 1). A total of 12 algorithms were used for modeling among six studies, while three of them were DL and the remaining were ML approaches (Fig. 2) (Table 2). DL models included neural networks (NN), ANN, and multilayer perceptron (MLP). ML approaches were logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), decision tree (DT), extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), and gradient-boosted decision trees (GBDT). The KNN was the best-performing algorithm with the highest average values for accuracy (0.92.6), sensitivity (0.833), specificity (1), and AUC (0.979) (Fig. 3) (Table 2). the input characteristics primarily comprised radiomics features with radiomics combined with clinical features, reported in six studies; in addition, genomic features were used in two studies.

Sensitivity and specificity

The pooled sensitivity was obtained at 0.87 [95% CI: 0.78–0.92], with mild heterogeneity noted with an I2 of 34.27 [95% CI: 0–94.3]. $\chi 2$ test of heterogeneity exhibited a Q of 7.61 (p-value = 0.18). The pooled specificity was demonstrated to be 0.86 [95% CI 0.67–0.95], with significant heterogeneity observed between the studies, represented by an I2 of 95.65 [95% CI: 93.45 −97.85]. The $\chi 2$ test yielded Q of 5 (p-value < 0.001) (Fig. 4).

Positive and negative diagnostic likelihood ratio (DLR)

The positive DLR found to be 6.32 [95% CI: 2.46–16.26], while showed significant heterogeneity with an I2 of 91.96 (95% CI: 91.96–97.56). The χ^2 test for heterogeneity deferred a Q of 95.42 (p-value < 0.001).

The pooled negative DLR was 0.16 [95% CI: 0.1–0.25], with moderate heterogeneity indicated by an I2 value of 44.6 [95% CI: 0–95.96]. The results of the $\chi 2$ test revealed a Q of 9.03 (p-value = 0.11) (Fig. 5).

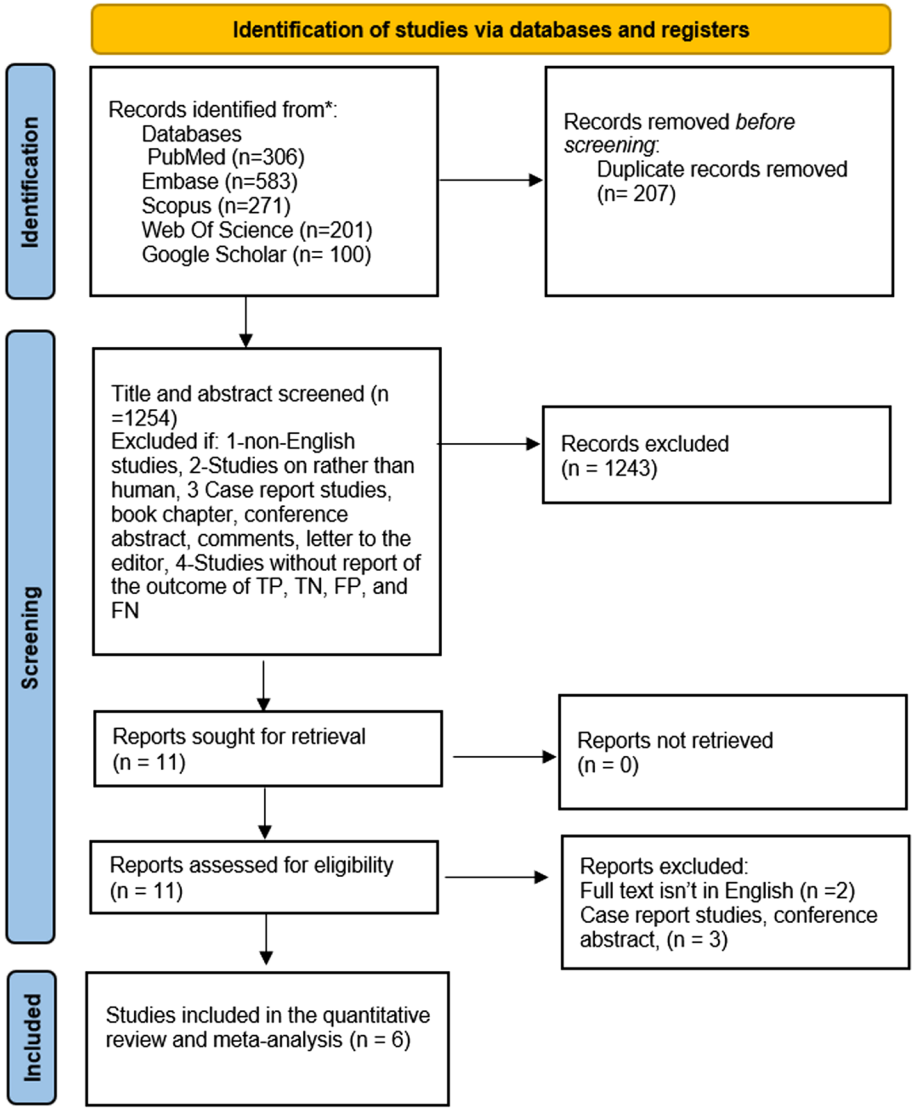

Fig. 1 PRISMA flowchart of the study selection process

Diagnostic score and diagnostic odds ratio

The diagnostic score was calculated as 3.7 [95% CI: 2.57–4.84], while demonstrating significant heterogeneity with an I2 of 99.05% [95% CI:95.75–99.34]. The $\chi 2$ test showed a Q (p-value < 0.001). The DOR was found to be 40.52 [95% CI: 13–126.27], with significant variation observed between the studies with the I2 of 100% [95% CI: 100–100]. The $\chi 2$ test revealed a Q of 4.3e +21(p-value < 0.001) (Fig. 6).

Performance metric (area under the curve, AUC)

The pooled performance of best-performing ML algorithm from each study in predicting recurrent PA was calculated as 0.89 [95% CI: 0.86–0.92] (Fig. 7).

Meta-analysis on all algorithms

Pooled sensitivity, specificity was 0.84 [95% CI: 0.79–0.87] and 0.92 [95% CI: 0.84–0.96] with negligible ($I^2 =$ 0%) and significant heterogeneity ($I^2 =$ 90.55%, respectively), respectively. The positive DLR was 10.21, and the negative DLR 0.18, corresponding to significant and very low heterogeneity ($I^2 =$ 83.28% and 0%), respectively. The diagnostic score was 4.04 and the DOR was 57.02 ($I^2 =$ 97.10% and 100%, respectively). The pooled area under the SROC curve had an AUC of 0.88 [95% CI: 0.85–0.90] (Supplementary file 2).

Table 1 Demographic and characteristics of inclusion studies

Author/Year	Type of study	Country	No of patients in train/test/validation	Recurrence criteria	Time to recurrence (months)	Type of tumor	Follow up (Months)	Mean age/female%	Inclusion criteria	Exclusion criteria	Type of treatment
Y. Liu et al/2019 [23]	Retrospective	China	354	Elevated morning serum cortisol level or 24-h urine free cortisol (24 hUFC) Lack of suppression of the morning cortisol level after a late evening dose of dexamethasone MRI confirmation of recurrence	28	ACTH-secreting	At least 12	34.0/82.2%	Patients with initial TSS for CD, at least 12-month follow-up, immediate remission post-TSS	Follow-up < 12 months, lack of follow-up, no immediate remission post-TSS	TSS
E. Y. Nadezhdina et al./2019 [24]	Retrospective	Russia	155/64	Increased evening salivary cortisol level No suppression of serum cortisol below 50 nmol/L (1.8 µg/dL) during the 1-mg dexamethasone suppression test Increased 24-h urine free cortisol level Increased concentrations and abnormal secretory rhythms of ACTH and cortisol Clinical recurrence of hypercortisolism	The 1-year recurrence rate was 4.7% The 3-year recurrence rate was 19.5%	ACTH-secreting	133.5	38/85.1%	Laboratory-confirmed early postoperative remission	Previous treatment (radiation therapy and/or neurosurgery)	TSS
L. F. Machado et al/2020 [25]	Retrospective	Brazil	27	Tumor growth sufficient to produce mass effect symptoms or pose a risk of future complications, particularly optic nerve compression Assessed through regular imaging studies, tumor volume tracking, and observation of mass effect symptoms	5.91 ±2.4 years after the first surgical approach	NFPA	70.92	50.4/66.7%	Clinically non-functioning pituitary macroadenoma At least two years of clinical and imaging follow-up documented MRI performed with the same equipment and protocol	Inadequate documentation of follow-up Lack of preoperative images with the desired MRI protocol Radiotherapy within two years of surgery for reasons other than tumor recurrence Second surgery within 2.5 years for reasons other than recurrence	Surgery, with some patients undergoing radiotherapy or additional surgery for recurrence

Table 1 (continued)

Author/Year	Type of study	Country	No of patients in train/test/validation	Recurrence criteria	Time to recurrence (months)	Type of tumor	Follow up (Months)	Mean age/female%	Inclusion criteria	Exclusion criteria	Type of treatment
Sh. Shahrestani et al/2021 [26]	Retrospective	USA	208/70/70	Evidence of tumor recurrence or progression, hormonal non-remission, or imaging/biochemical evidence of recurrence or progression following surgery	NA	GH-secreting (191) + ACTH-secreting (146) + Mammosomatotroph (11)	68.2	41.7/63.8%	Patients undergoing endonasal transsphenoidal resection of FPAs, including GH-secreting, ACTH-secreting, and mammosomatotroph adenomas	Non-functioning/silent adenomas, non-transsphenoidal tumor resections	Gross-total resection (87.6%), subtotal resection (12.4%)
Ch. Shen et al/2023 [27]	Retrospective	China	80/34	Residual tumor regrowth defined as an increase in the maximum tumor diameter > 2 mm in any direction on MRI, starting from the day of surgery to the follow-up endpoint	NA	NFPA	24–108 (mean 64.7)	51.7/39.62%	NF-PitNET patients with postoperative tumor residue > 10 mm, ≥ 2 years of follow-up, and MRI using the same protocol	Inadequate documentation, incomplete imaging, RT or second surgery within 2 years	surgery
J. Zhong et al/2024 [28]	Retrospective	China	126	Rumor regrowth after GTR, identified during the follow-up period when MRI reveals the appearance of a new mass	56 months (recurrence), 19.5 months (progression)	NFPA	94.5 (28–153)	50.3/0%	Clinical and pathological diagnosis of NFPA, > 18 years, complete medical and imaging records, 3.0 T MRI examination, WHO 2017 criteria compliance	Non-microscopic surgery, history of craniofacial/sellar surgery, incomplete data or lost follow-up, prior radiotherapy/medication, concurrent intracranial/malignant tumors	MTSS (GTR and STR)

Abbreviations: *NA* Not available, *TSS* Transsphenoidal surgery, *CD* Cushing's disease, *MRI* Magnetic resonance imaging, *24 hUFC* 24-h Urine free cortisol, *NFPA* Non-functional pituitary adenoma, *ACTH* Adrenocorticotropic hormone, *FPAs* Functional pituitary adenomas, *NFPA* non-functional pituitary adenoma, *GH* Growth hormone, *GTR* Gross total resection, *STR* Subtotal resection

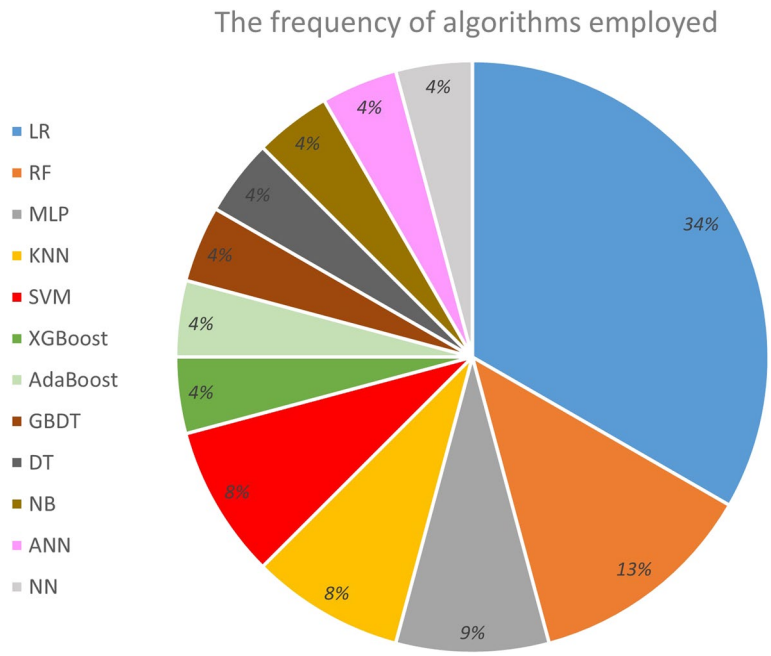

Fig. 2 Frequency of algorithms used in the analyzed studies

Sub group between LR-based and non-LR-based algorithms

Compared to the non-LR-based algorithm, the LR-based algorithm exhibited a slightly higher AUC (0.89 [95% CI: 0.86–0.92] vs. 0.87 [95% CI: 0.84–0.90]) and sensitivity (0.86 [95% CI: 0.80–0.90] vs 0.81 [95% CI: 0.74–0.87]), but its specificity was lower (0.81 [95% CI: 0.71–0.88] vs. 0.96 [95% CI: 0.87–0.99]). While LR-based algorithms had a slightly higher AUC and sensitivity, their specificity was lower, indicating that non-LR-based algorithms may better rule out non-recurrence. The LR-based algorithm also showed a lower positive likelihood ratio (4.4 [95% CI: 2.9–6.8] vs. 22.1 [95% CI: 6–81.2]) and a comparable negative likelihood ratio (0.17 [95% CI: 0.12–0.25] vs. 0.19 [95% CI: 0.13–0.28]). Additionally, although the DOR for the non-LR-based algorithm (26 [95% CI: 14–48]) appeared numerically higher than the LR-based algorithm (115 [95% CI: 29–449]), this difference was not statistically significant (*P*-value: 0.15), indicating no strong evidence to suggest a meaningful difference in DOR between the two algorithms.

Bias assessment (PROBAST)

Among the four evaluated domains, including participants, predictors, outcomes, and analysis, 82.02% of studies demonstrated a low risk of bias. In contrast, 10.8% of studies showed high risk, mainly related to predictors and outcomes. No studies were identified with unclear bias (Supplementary file 3, Fig. S1).

Applicability assessment (PROBAST)

Regarding the applicability, 90% of studies had low concerns across participants, predictors, outcomes, and analysis. However, 10% showed high concerns, particularly in the predictors category. None of the studies were rated as unclear in terms of applicability (Supplementary file 3, Fig. S2).

Discussion

This systematic review and meta-analysis aimed to assess how accurately ML models can predict PA recurrence following surgery. In this study, two types of meta-analysis, one based on the best-performing model of each study and the other according to the data of all ML models (all-models approach) was performed. The results of meta-analysis based on best-predictor model of each study (RF, ANN, NN and LR), demonstrated a pooled sensitivity of 0.87 shows a high ability to correctly identify cases of recurrence, while the pooled specificity of 0.86 demonstrates their ability to exclude cases of non-recurrence.

In addition, based best models, the effectiveness of the diagnostic models is supported by a positive DLR of 6.32, indicating that when the test result is positive, the probability of recurrence will considerably increase. On

Table 2 ML models characteristics and performance metrics

Author/Year	Validation	Type of reference	Input characteristics	Selected features	Method of radiomics	No. of extracted/final (features)	ML algorithm	Best predictor	Accuracy	Sensitivity(Recall)	Specificity	Precision	AUROC
Y. Liu et al/2019 [23]	fivefold cross-validation	MRI	Radiomics features and clinical features	Preoperative age, gender, BMI, disease course, tumor size, Knosp grade, endocrinological results; postoperative pathological and IHC results, 7-day endocrinological results	Automated	17	DT, RF, GBDT, AdaBoost, XGBoost, LR, NB	RF	NA	0.87	0.581	NA	0.779
E. Y. Nadezhdina et al./2019 [24]	NA	MRI	Radiomics features and clinical features	Age Sex Duration of disease (months) Type of adenoma Morning postoperative ACTH level Morning postoperative cortisol level	Automated	6	LR, ANN	ANN	0.92	0.75	0.97	NA	0.912
L. F. Machado et al/2020 [25]	threefold cross-validation	MRI	Radiomics features and clinical features	MRI images, age-at-first-surgery, gender, remnant lesion presence after the first surger	Automated	16 (single-slice 2D), 9 (multi-slice 2D)	3D Radiomics Features and 2D Radiomics Features KNN, RF, LR, SVM and MLP	RF (3D Radiomics features)	0.963	0.917	1	NA	0.962
Sh. Shahrestani et al/2021 [26]	tenfold cross-validation	MRI	Radiomics features and clinical features	Patient demographics, clinical symptoms, preoperative exam findings, tumor characteristics, extent of resection, CSF leak status, perioperative complications, hormonal remission, recurrence/progression outcomes	Automated	11	NN	NN	0.871	0.895	0.769	0.944	0.917

Table 2 (continued)

Author/Year	Validation	Type of reference	Input characteristics	Selected features	Method of radiomics	No. of extracted/final (features)	ML algorithm	Best predictor	Accuracy	Sensitivity(Recall)	Specificity	Precision	AUROC
Ch. Shen et al/2023 [27]	tenfold cross-validation	MRI	Radiomics features and clinical features and genomic features	Clinical, radiological, and pathological data (e.g., sex, age, BMI, headache, vision changes, Knosp classification, cystic transformation, Hardy classification, tumor consistency, ki-67 index, etc.)	Automated	854	Combined clinical and radiomic LR, clinical risk factors LR, radiomic features LR	LR (clinical and radiomic features)	0.888	0.848	0.933	NA	0.929
J. Zhong et al/2024 [28]	tenfold cross-validation	MRI	Radiomics features and clinical features and genomic features	Age, tumor size, modified Knosp grade, Ki-67 index, tumor type, resection extent, hormonal levels, clinical symptoms, hypopituitarism status	Automated	17	LR	LR	NA	0.92	0.723	NA	0.9

Abbreviations: *NA* Not available, *MRI* Magnetic resonance imaging, *IHC* Immunohistochemistry, *DT* Decision tree, *RF* Random forest, *GBDT* Gradient boosting decision Tree, *LR* Logistic regression, *NB* Naive bayes, *ANN* Artificial neural network, *KNN* K-nearest neighbors, *SVM* Support vector machine, *MLP* Multilayer perceptron, *NN* Neural network, *CSF* Cerebrospinal fluid, *BMI* Body mass index, *AUROC* Area under the receiver operating characteristic Curve

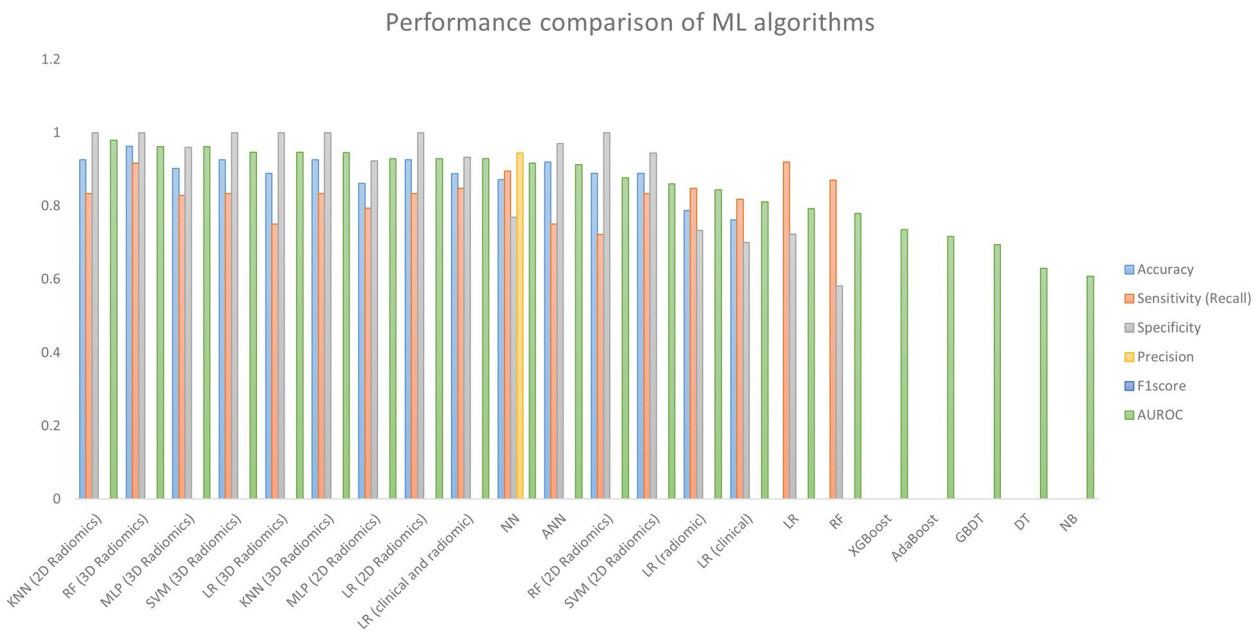

Fig. 3 Performance comparison of ML algorithms, including Accuracy, Sensitivity, Specificity, F1 Score, and AUROC

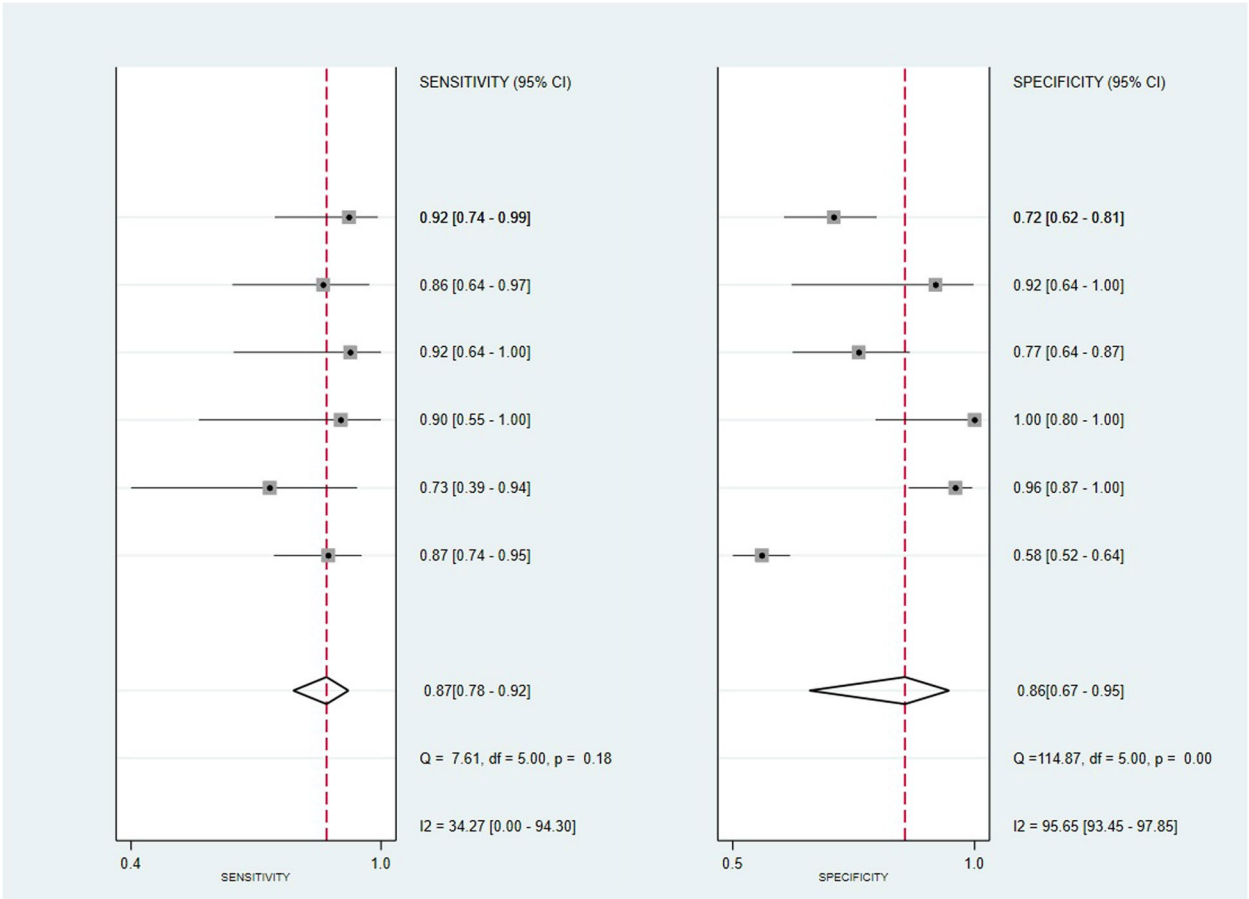

Fig. 4 Sensitivity and specificity of ML algorithms

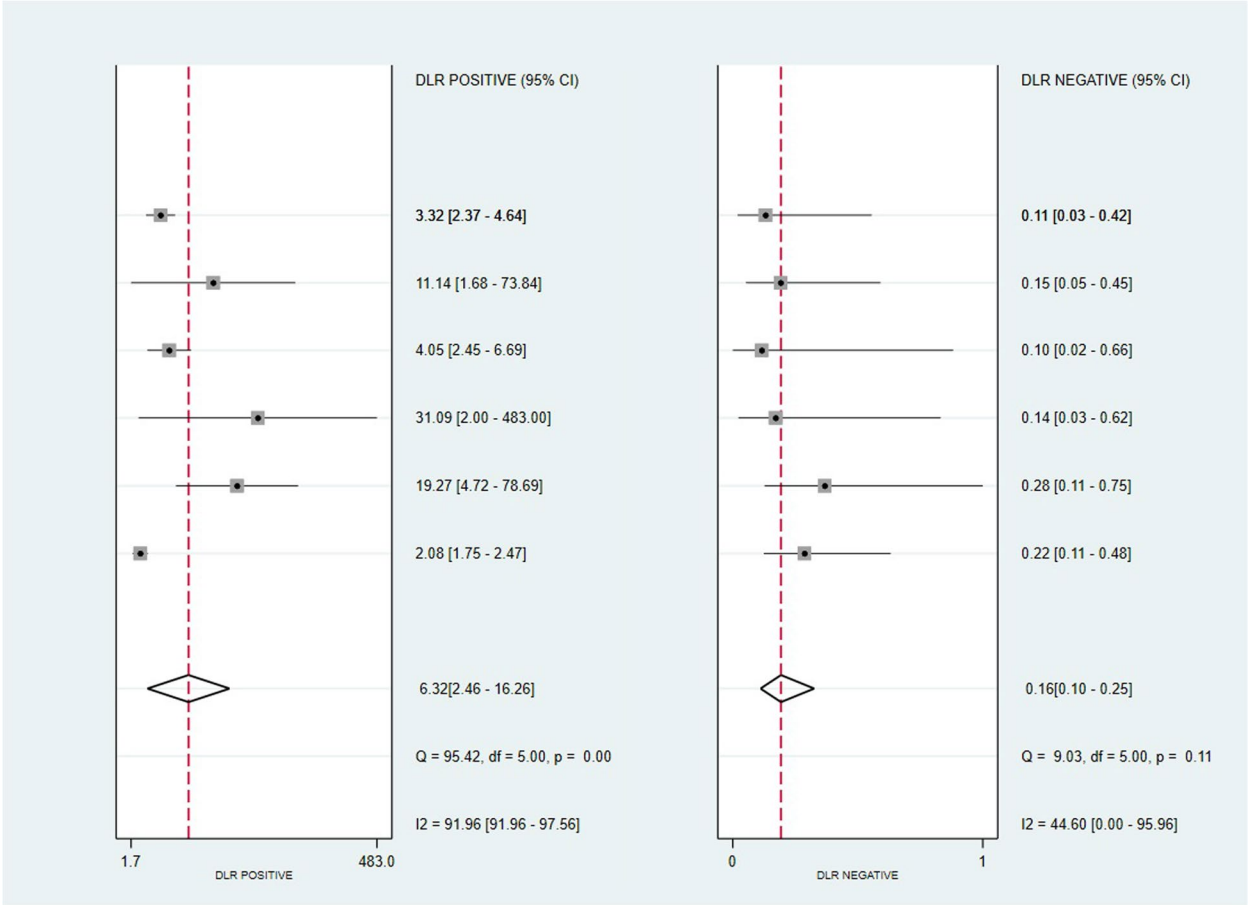

Fig. 5 Positive and negative DLR of ML algorithms

the other hand, the negative DLR of 0.16 depicts the decreasing recurrence probability when the test result is negative. The DOR of 40.52 shows how the models can distinguish between cases that return and those that do not. The AUC of 0.89 shows that ML is fairly accurate overall and confirms that they may be powerful tools for making predictions. These findings suggest that ML-based models may be reliable for predicting recurrence and guiding postoperative management strategies.

When all models and algorithms were included in the meta-analysis, the pooled sensitivity was 0.83, the specificity was 0.95, AUC of 0.88 and the DOR was 83.18, This approach demonstrated higher specificity and DOR but slightly lower sensitivity compared to the meta-analysis of the best-performing algorithms. However, the substantial heterogeneity observed in the all-model meta-analysis underscores the need for further refinement and standardization of ML models to ensure consistent reliability across diverse datasets and clinical contexts. The subgroup analysis showed that LR-based algorithms had slightly higher AUC (0.88 vs. 0.87), sensitivity (85% vs. 81%), and specificity (89% vs. 96%). In contrast, non-LR-based algorithms exhibited a significantly higher DOR (115 vs. 83.18) and positive likelihood ratio (22.1 vs. 7.6). These results illustrate that LR-based algorithms provide good diagnostic performance, which would help in many clinical settings. On the other hand, non-LR-based algorithms are better at confirming positive cases, which could be especially useful when confirmation is very important.

Comparison of ML algorithms

Y. Liu and colleagues conducted a study to predict the chances of recurrence after TSS for CD [15]. They used seven machine learning models based on 17 factors and compared them with two traditional models (LR and Naive Bayes). The models showed moderate accuracy, with the Random Forest model achieving the highest AUC of 0.78, followed by XGBoost, AdaBoost, and GBDT, which outperformed LR, DT, and Naive Bayes. Nine key predictors were identified, with the top three

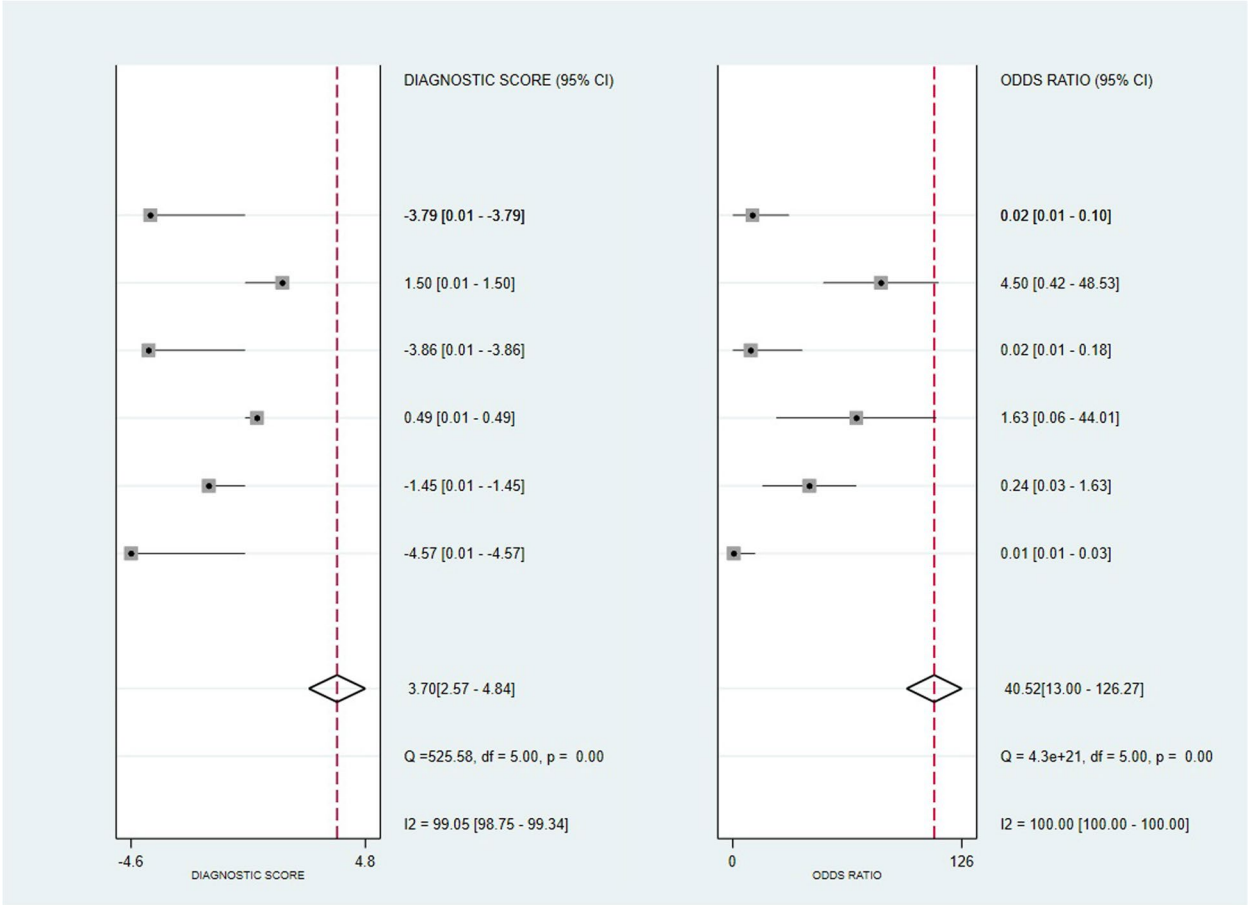

Fig. 6 Diagnostic score and diagnostic odds ratio of ML algorithms

being age, the lowest morning serum cortisol after surgery, and morning ACTH levels. They tested different numbers of variables with the seven algorithms to see how it affected performance. Most models reached a stable performance with nine variables. Adding more variables improved AdaBoost slightly, but the performance of Naive Bayes and LR decreased. Among the models, DT, LR and NB are interpretable models. At the same time, GBDT, RF, AdaBoost, and XGBoost are unexplainable, in which the function between the variables and the response is invisible to the user. However, the contribution of each variable could be inferred according to the feature selection method, and univariate analysis might indicate the direction.

An LR-based nomogram model developed by J. Zhong et al. effectively predicted postoperative recurrence after TSS for NFPAs in men, demonstrating a strong predictive capability with an AUC of 0.9 [29]. Internal validation confirmed that the model exhibited excellent discrimination and calibration. Female patients were excluded from this study due to a higher prevalence of conditions such as type 2 diabetes mellitus, myocardial infarction, cerebral infarction, and fractures, which may impact postoperative outcomes and mortality risk factors. The predictive model ultimately included three key predictors identified through least absolute shrinkage and selection operator (LASSO): Ki67, Modified Knosp grade, and resection extent.

Neural network (NN) models have proven highly effective in predicting adenoma recurrence. In a study by Sh. Shahrestani et al., a multilayered NN, analyzed outcomes in 348 patients with FPA as causing CD or acromegaly [30]. Key variables, including MRI follow-up assessments, diabetes insipidus (DI), tumor traits, endocrine test results, prior craniotomies, and hospital types, were input into the NN. The model for the entire cohort outperformed those specific to CD and acromegaly, achieving an AUC of 0.91. This indicates that diverse samples of FPA patients, given adequate statistical power, can yield highly accurate predictions for postoperative outcomes. A three-layer ANN model based on age, disease duration, MRI data on adenoma,

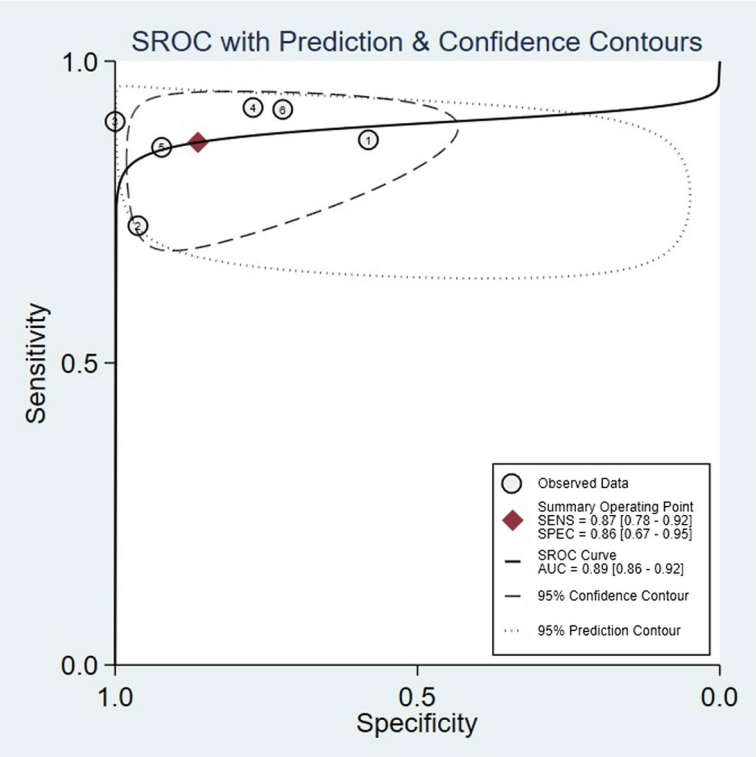

Fig. 7 Summary receiver operator characteristic curve (SROC) of ML algorithms

morning postoperative levels of ACTH and cortisol suggested by E. Y. Nadezhdina et al. [31]. This model demonstrated an accuracy of 84% in the validation sample when predicting the three-year recurrence rate for patients with CD who have undergone ETS. ROC analysis showed the high efficacy of the model with AUC 0.912. The model performs over-diagnosis in 15% of cases (i.e., predicts recurrence in 41 when there is actual remission in 6 of them) and under-diagnosis in 7% of cases (i.e., predicts remission in 178 when there is actual recurrence in 12 of them). So, over-diagnosis is slightly higher than under-diagnosis. Such a strategy is consistent with a high level of alertness for recurrence.

Radiomic prediction of pituitary adenoma recurrence

Radiomics, predominantly used in oncology, is an approach to quantitative perception of medical imaging, which operates on the assumption that biomedical images hold disease-specific insights beyond human visual perception and traditional inspection. The process involves four key steps: segmentation, processing, feature extraction, and feature selection or dimension reduction [32]. Machado et al., evaluated the prognostic power of preoperative MRI radiomics features based on two extraction modalities (2D and 3D), combined with ML models to differentiate the recurrent NFPA patients [1]. Six and thirteen radiomic feature found to be statistically different in recurrent and non-recurrent lesions for 2D and 3D radiomics, respectively. The majority of the features (2D and 3D) were related to energy, total-energy, and non-uniformity, which cannot be seen or interpreted with the naked eye because they are mostly obtained from filtered images. While the ML experiments trained with both 2D and 3D radiomics feature sets produced excellent results (AUC >0.85), ML models trained with 3D features achieved superior accuracies than when trained with 2D features and used fewer features.

In another study to predict the recurrence of NFAPs using the radiomics feature, Shen et al. employed both pre-and postoperative MRI radiomics features to compare the models based on T1-weighted (T1 WI), T2-weighted (T2 WI) and contrast-enhanced T1 (T1 CE) sequences to differentiate regrowth and non-regrowth groups [7]. The best predictive power was obtained when pre- and postoperative radiomic features were combined rather than using single pre- or postoperative images. This indicates that the postoperative images are significant referents for investigating residual tumors. Integrating two independent clinical variables, including Knosp classification (grades 3 or 4) and preoperative tumor volume doubling time (TVDT) with radiomics based on the three modalities (T1 WI, T1 CE, T2 WI) using the LR

approach resulted in an excellent predictive power (AUC of 0.92 in training cohort and 0.88 in the test set) and performed significantly better than the single clinical or radiomic model in both the training and test sets.

Limitation

Due to the retrospective single-center design of most studies, bias is inevitable to some extent, and a large amount of missing data is possible. Several studies lacked clearly defined test and training groups, which is crucial in AI studies. In these cases, the absence of specified cases and controls for each patient group made it impossible to calculate key metrics such as TP, TN, FP, and FN. Additionally, patients developing suboptimal outcomes may not have been adequately represented due to improper sample sizes and short follow-up periods. The high heterogeneity in specificity ($I^2 = 95.65\%$) indicates that the models' ability to rule out non-recurrence varies widely, possibly due to differences in study populations or imaging protocols. Another major source of heterogeneity is the diversity of machine learning algorithms used in the studies, as each algorithm has distinct structures and functionalities, which can contribute to variations in performance.

Future direction

To improve the applicability and performance of ML models in predicting recurrence in PA, future original studies should focus on expanding the dataset through multi-center, prospective research with larger and more diverse patient populations [14]. A more robust study design, including well-defined test and training groups, adequate sample sizes, and longer follow-up periods, is essential to accurately calculate key metrics such as TP, TN, FP, and FN. Standardizing imaging protocols, particularly MRI parameters. Additionally, optimization and comparison of various ML algorithms, including DL models, should be explored further to enhance predictive accuracy, particularly for different PA subtypes (e.g., functional vs. non-functional). Finally, external validation of predictive models in independent multi-center cohorts is necessary to assess their real-world applicability and ensure that they generalize well across diverse patient populations.

Conclusion

AI-based models show potential in predicting recurrence in patients with both FPAs and NFPAs. These models demonstrate a discriminatory ability exceeding 80%. Various variables, including clinical, imaging, and radiomics features, have been identified as predictors, highlighting the capability of AI models to process diverse data for predictions. The pooled results from the best-performing and all-model approaches suggest promising predictive power, offering the potential to reduce bias in clinical decision-making. However, a significant limitation of the current models is the lack of external validation, coupled with the small size of the training cohorts. Additionally, high heterogeneity in specificity across studies underscores the need for further validation. To bridge the gap between research and clinical application, it is essential to conduct rigorous validation of these models using multi-center, prospective datasets with standardized imaging protocols.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12902-025-01955-8.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements
Not applicable.

Clinical trial number
Not applicable.

Authors' contributions
The conception and design of the study: Ibrahim Mohammadzadeh (IM), Behnaz Niroomand (BN), Hamid Borghei-Razavi (HBR) and Bardia Hajikarimloo (BH). Acquisition of data: IM, Pooya Eini and BN. Analysis and interpretation of data: IM, BH, and BN. Drafting the article: IM, BH, Nasira Faizi and Mohammad Amin Habibi (MAH). Revising it critically for important intellectual content: MAH, Mohammadmahdi Sabahi, Alireza Mohseni, Michael Karsy and Abdulrahman Albakr. Final approval of the version to be submitted: All authors.

Funding
The authors did not receive support from any organization for the submitted work.

Data availability
The datasets and materials used and analyzed during the current study are available from the corresponding author, I. Mohammadzadeh, upon reasonable request.

Declarations

Ethics approval and consent to participate
The study is deemed exempt from receiving ethical approval.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Skull Base Research Center, Loghman-Hakim Hospital,, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Department of Neurological Surgery, University of Virginia, Charlottesville, VA, USA. [3]School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [4]Toxicological Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

[5]Department of Neurosurgery, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran. [6]Clinical Research Development Unit of Torfe Medical Center, Shahid Beheshti University of Medical Science, Tehran, Iran. [7]Department of Neurological Surgery, Pauline Braathen Neurological Center, Cleveland Clinic Florida, Weston, FL, USA. [8]Department of Surgery, Division of Neurosurgery, King Saud University, Riyadh, Saudi Arabia. [9]Department of Neurosurgery, University of Michigan, Ann Arbor, MI, USA. [10]Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, USA. [11]Neurological Surgery at CCLCM of CWRU Director of Minimally Invasive Cranial and Pituitary Surgery Program Research Director, Neuroscience Institute, Cleveland Clinic Florida Region, Cleveland, OH, USA. [12]Department of Skull Base Research Center, Loghman-Hakim Hospital, Shahid Beheshti University of Medical Sciences, Skull-Base Neurosurgery From Shahid Beheshti Medical University, Tehran, Iran.

Received: 1 March 2025 Accepted: 9 May 2025
Published online: 01 July 2025

References

1. Machado LF, Elias PCL, Moreira AC, Dos Santos AC, Murta Junior LO. MRI radiomics for the prediction of recurrence in patients with clinically non-functioning pituitary macroadenomas. Comput Biol Med. 2020;124:516-524. https://doi.org/10.1016/j.compbiomed.2020.103966. (In eng).
2. Molitch ME. Diagnosis and treatment of pituitary adenomas: a review. JAMA. 2017;317(5):516–24.
3. Melmed S, Kaiser UB, Lopes MB, et al. Clinical biology of the pituitary adenoma. Endocr Rev. 2022;43(6):1003–37. https://doi.org/10.1210/endrev/bnac010. (In eng).
4. Marrero-Rodríguez D, Vela-Patiño S, Martinez-Mendoza F, et al. Genomics, transcriptomics, and epigenetics of sporadic pituitary tumors. Arch Med Res. 2023;54(8):102915. https://doi.org/10.1016/j.arcmed.2023.102915. (In eng).
5. Jimenez MA, Horowitz MA, Gendreau JL, et al. Characterizing Disparities in Access to Surgery for Pituitary Adenomas: A National Cancer Database Analysis. J Clin Endocrinol Metab. 2025:dgaf212. https://doi.org/10.1210/clinem/dgaf212.
6. Tritos NA, Miller KK. Diagnosis and management of pituitary adenomas: a review. Jama. 2023;329(16):1386–98. https://doi.org/10.1001/jama.2023.5444. (In eng).
7. Shen C, Liu X, Jin J, et al. A Novel Magnetic Resonance Imaging-Based Radiomics and Clinical Predictive Model for the Regrowth of Postoperative Residual Tumor in Non-Functioning Pituitary Neuroendocrine Tumor. Medicina. 2023;59(9):1525. https://www.mdpi.com/1648-9144/59/9/1525.
8. Lu L, Wan X, Xu Y, Chen J, Shu K, Lei T. Prognostic factors for recurrence in pituitary adenomas: recent progress and future directions. Diagn. 2022;12(4):977 https://www.mdpi.com/2075-4418/12/4/977.
9. Chukwujindu E, Faiz H, Ai-Douri S, Faiz K, De Sequeira A. Role of artificial intelligence in brain tumour imaging. Eur J Radiol. 2024;176:111509. https://doi.org/10.1016/j.ejrad.2024.111509. (In eng).
10. Mohammadzadeh I, Niroomand B, Eini P, Khaledian H, Choubineh T, Luzzi S. Leveraging machine learning algorithms to forecast delayed cerebral ischemia following subarachnoid hemorrhage: a systematic review and meta-analysis of 5,115 participants. Neurosurg Rev. 2025;48(1):26.
11. Mohammadzadeh I, Niroomand B, Hajikarimloo B, et al. Can we rely on machine learning algorithms as a trustworthy predictor for recurrence in high-grade glioma? A systematic review and meta-analysis. Clin Neurol Neurosurg. 2025;249:108762. https://doi.org/10.1016/j.clineuro.2025.108762.
12. Hajikarimloo B, Mohammadzadeh I, Nazari MA, et al. Prediction of facial nerve outcomes after surgery for vestibular schwannoma using machine learning-based models: a systematic review and meta-analysis. Neurosurg Rev. 2025;48(1):79. https://doi.org/10.1007/s10143-025-03230-9.
13. Mohammadzadeh I, Hajikarimloo B, Niroomand B, et al. Application of artificial intelligence in forecasting survival in high-grade glioma: systematic review and meta-analysis involving 79,638 participants. Neurosurg Rev. 2025;48(1):240. https://doi.org/10.1007/s10143-025-03419-y.
14. Zhang W, Wu X, Wang H, et al. Federated learning for predicting postoperative remission of patients with acromegaly: a multicentered study. World Neurosurg. 2025;193:1036–46. https://doi.org/10.1016/j.wneu.2024.10.091. (In eng).
15. Liu Y, Liu X, Hong X, et al. Prediction of recurrence after transsphenoidal surgery for cushing's disease: the use of machine learning algorithms. Neuroendocrinology. 2019;108(3):201–10. https://doi.org/10.1159/000496753. (Article) (In English).
16. Mohammadzadeh I, Niroomand B, Shahnazian Z, et al. Machine learning for predicting poor outcomes in aneurysmal subarachnoid hemorrhage: A systematic review and meta-analysis involving 8445 participants. Clin Neurol Neurosurg. 2025;249:108668. https://doi.org/10.1016/j.clineuro.2024.108668.
17. Dai C, Sun B, Wang R, Kang J. The application of artificial intelligence and machine learning in pituitary adenomas. Front Oncol. 2021;11:784819.
18. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial Intelligence in cancer research and precision medicine. Cancer Discov. 2021;11(4):900–15. https://doi.org/10.1158/2159-8290.Cd-21-0090. (In eng).
19. Bioletto F, Prencipe N, Berton AM, et al. Radiomic Analysis in Pituitary Tumors: Current Knowledge and Future Perspectives. J Clin Med. 2024;13(2):336 https://www.mdpi.com/2077-0383/13/2/336..
20. Zheng B, Zhao Z, Zheng P, et al. The current state of MRI-based radiomics in pituitary adenoma: promising but challenging. Front Endocrinol (Lausanne). 2024;15:1426781. https://doi.org/10.3389/fendo.2024.1426781. (In eng).
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71. (In eng).
22. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. https://doi.org/10.7326/m18-1376. (In eng).
23. Liu Y, Liu X, Hong X, et al. Prediction of recurrence after transsphenoidal surgery for Cushing's disease: the use of machine learning algorithms. Neuroendocrinology. 2019;108(3):201–10.
24. Nadezhdina EY, Rebrova OY, Grigoriev AY, et al. Prediction of recurrence and remission within 3 years in patients with Cushing disease after successful transnasal adenomectomy. Pituitary. 2019;22(6):574–80.
25. Machado LF, Elias PC, Moreira AC, Dos Santos AC, Junior LOM. MRI radiomics for the prediction of recurrence in patients with clinically non-functioning pituitary macroadenomas. Comput Biol Med. 2020;124:103966.
26. Shahrestani S, Cardinal T, Micko A, et al. Neural network modeling for prediction of recurrence, progression, and hormonal non-remission in patients following resection of functional pituitary adenomas. Pituitary. 2021;24:523–9.
27. Shen C, Liu X, Jin J, et al. A Novel Magnetic Resonance Imaging-Based Radiomics and Clinical Predictive Model for the Regrowth of Postoperative Residual Tumor in Non-Functioning Pituitary Neuroendocrine Tumor. Medicina. 2023;59(9):1525.
28. Zhong J, Chen Y, Wang M, et al. Risk factor analysis and prediction model to establish recurrence or progression of non-functioning pituitary adenomas in men after transnasal sphenoidal surgery. Sci Rep. 2024;14(1):21607.
29. Zhong J, Chen Y, Wang M, et al. Risk factor analysis and prediction model to establish recurrence or progression of non-functioning pituitary adenomas in men after transnasal sphenoidal surgery. Sci Rep. 2024;14(1):21607. https://doi.org/10.1038/s41598-024-72944-5. (In eng).
30. Shahrestani S, Cardinal T, Micko A, et al. Neural network modeling for prediction of recurrence, progression, and hormonal non-remission in patients following resection of functional pituitary adenomas. Pituitary. 2021;24(4):523–9. https://doi.org/10.1007/s11102-021-01128-5. (In eng).
31. Nadezhdina EY, Rebrova OY, Grigoriev AY, et al. Prediction of recurrence and remission within 3 years in patients with cushing disease after successful transnasal adenomectomy. Pituitary. 2019;22(6):574–80. https://doi.org/10.1007/s11102-019-00985-5. (In eng).
32. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging. 2020;11(1):91. https://doi.org/10.1186/s13244-020-00887-2.

Publisher's Note