npj Digital Medicine

https://doi.org/10.1038/s41746-025-02189-7

Article in Press

An MRI radiomics approach using invasion-based weak supervision for identifying and evaluating aggressive PitNETs

Received: 28 August 2025

Accepted: 18 November 2025

Cite this article as: Wang, Y., Guan, X., Ma, S. *et al.* An MRI radiomics approach using invasion-based weak supervision for identifying and evaluating aggressive PitNETs. *npj Digit. Med.* (2025). https://doi.org/10.1038/s41746-025-02189-7

Yangyang Wang, Xiudong Guan, Shunchang Ma, Yuyan Zhang, Jun Yang, Yan Liu, Yueqian Sun, Li Ma, Deling Li, Yongqiang Tang, Chuanbao Zhang & Wang Jia

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

An MRI radiomics approach using invasion-based weak supervision for identifying and evaluating aggressive PitNETs

Yangyang Wang[1], Xiudong Guan[1], Shunchang Ma[2], Yuyan Zhang[3], Jun Yang[4], Yan Liu[1], Yueqian Sun[5], Li Ma[1], Deling Li[1], Yongqiang Tang[6*], Chuanbao Zhang[1*], Wang Jia[1,2,7*]

[1]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China.

[2]Beijing Neurosurgery Research Institute, Capital Medical University, Beijing, China.

[3]Department of Neurosurgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China.

[4]Department of Neurosurgery, Beijing Luhe Hospital, Capital Medical University, Beijing, China.

[5]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China.

ARTICLE IN PRESS

ARTICLE IN PRESS

[6]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

[7]China National Clinical Research Center for Neurological Diseases (NCRC-ND), Beijing, China.

*Correspondence

Yongqiang Tang, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (yongqiang.tang@ia.ac.cn)

Chuanbao Zhang, Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China (chuanbao123@139.com)

Wang Jia, Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China (jwjsttyy@163.com).

ARTICLE IN PRESS

ARTICLE IN PRESS

Abstract

Pituitary neuroendocrine tumor (PitNET) aggressiveness critically affects treatment and prognosis, yet reliable noninvasive preoperative tools remain lacking. We developed a deep learning radiomics (DLR) model integrating automatic segmentation, feature extraction, selection, and DLR score computation, trained on the training cohort and validated on the remaining cohorts (total n=1,089 from three medical centers). Using nnUnet and a fine-tuned Swin Transformer, 13 key features were identified to construct the model. The DLR score demonstrated strong correlation with Knosp and Hardy-Wilson invasion classifications, while outperforming them in predicting recurrence and indicating aggressive pathological markers (Ki-67, p53, macrophages) and revealing biological pathways (MAPK, TGF-β). The model was further implemented into an online platform, enabling clinical deployment. This noninvasive preoperative approach provides a robust imaging biomarker for identifying and evaluating PitNET aggressiveness and may support individualized treatment strategies.

Keywords: aggressive behaviors; pituitary neuroendocrine tumors; deep learning; radiomics; MRI; multi-omics

ARTICLE IN PRESS

ARTICLE IN PRESS

Introduction

Pituitary neuroendocrine tumor (PitNET), also known as pituitary adenoma, accounts for 10%-20% of all intracranial tumors, with a reported prevalence of approximately 70-100 cases per 100,000 individuals[1,2]. Although PitNET is considered benign, over 40% of cases exhibit an invasive tumor during development[3]. The treatment of invasive PitNET is challenging due to difficulties related to complete surgical resection, which has contributed to rapid recurrence and the need for multiple-line treatments[4].

Knosp and Hardy-Wilson (Hardy) classifications were developed last century to assess PitNET invasion by region (lateral or superior/inferior) **(Fig. 1a)**, and guide clinical management[5-8], but have limited biological relevance[9-11]. Recently, the WHO and European Society of Endocrinology (ESE) introduced the concept of aggressive pituitary tumors (APTs), characterized by invasion, rapid growth, recurrence, pathological changes (Ki-67, mitotic count, p53), and treatment resistance[12,13].

Aggressive pituitary tumors (APTs) show multiply recurrent and treatment refractoriness than non-aggressive lesions, often necessitating personalized management. Preoperative risk identification can guide early referral to high-volume pituitary centers, selection of multimodal therapy, and intensified surveillance[4]. In refractory or multiply recurrent APTs, postoperative management often requires combined-modality therapy—including

ARTICLE IN PRESS

radiotherapy, alkylating chemotherapy such as temozolomide (TMZ), and immunotherapy[14,15]. Accordingly, early identification and evaluation of aggressive tumor behavior can support more precise treatment selection and guide clinical decision-making.

Current studies on aggressive behaviors mostly rely on surgical specimens to examine molecular markers via immunohistochemistry (IHC) or second-generation sequencing[16]. Yet no effective tool or widely accepted biomarker exists for early identification and evaluation. Specimen-based methods preclude preoperative evaluation of tumor aggressiveness, hindering personalized treatment planning. Thus, developing broadly applicable preoperative approaches for identifying and evaluating aggressive behaviors is essential[17].

Recent key algorithmic breakthroughs have markedly advanced MRI data processing and the extraction of more integrated information[18,19], thereby enabling downstream tasks based on MRI-derived features, including the evaluation of therapeutic response[20,21], tumor recurrence[22], and molecular markers[23,24] in brain tumors. In particular, MRI-based studies also feature non-invasiveness prior to surgery—this characteristic not only offers convenience for operation and ease of application, but also alleviates the economic burden on patients to a certain extent.

ARTICLE IN PRESS

Herein, a large number of PitNETs with multi-omics data, such as MRI, IHC, and transcriptome data, were collected. We applied a deep learning approach based on MRI images to identify and evaluate the aggressiveness of PitNETs, including correlating with Knosp and Hardy invasion classification, predicting tumor recurrence, and indicating aggressive pathology markers as well as revealing aggressive biological processes (BP) **(Fig. 1b and 1c)**.

Results

Patient characteristics

In total, 1089 PitNETs were enrolled in this study. The detailed workflow of cohort collection and exclusion is summarized in **Fig. 1b**. **Table 1** presents the characteristics of the PitNETs in the training cohort, independent test cohort, external test cohort and prospective test cohort. Age, gender, Knosp and Hardy classifications showed no significant differences across cohorts, while initial symptoms, surgical approach and IHC characteristics significantly varied.

Deep learning radiomics model construction

After tumor segmentation and fine-tuning of the deep network for feature extraction (detailed results are provided in **Supplementary Tables 1-2** and **Supplementary Figures**

ARTICLE IN PRESS

1-2), a total of 2304 (768×3) deep learning features (DLFs) were extracted from three regions of interest (ROIs) per patient (**Fig. 1d**). After feature selection, and the deep learning radiomics (DLR) model was established by weighting each selected feature using its estimated regression coefficient in the multivariable logistic regression analysis. The resulting DLR model was constructed by 13 DLFs (denoted as f1-f13). The detailed screening results of DLFs are provided in **Supplementary Data 1.** The DLR score was calculated as follows: DLR score = (f1 × 1.35) + (f2 × 0.03) + (f3 × -0.54) + (f4 × -0.40) + (f5 × 0.71) + (f6 × -0.59) + (f7 × -0.13) + (f8 × -0.26) + (f9 × 0.77) + (f10 × -0.90) + (f11 × 1.42) + (f12 × -0.01) +(f13 × -0.53).

DLR model correlates with PitNET invasion

The Spearman rho correlation coefficient between the DLR score and Knosp classification was 0.640, 0.566, 0.632, 0.725 in the training cohort, independent test cohort, external test cohort and prospective test cohort **(Fig. 2a)**, respectively. Similarly, the Kendall tau correlation coefficient between the DLR score and Hardy classification was 0.387, 0.344, 0.343, 0.491 in these four cohorts (**Fig. 2b**). These results demonstrated significant positive correlations between the DLR score and Knosp or Hardy classification in the four cohorts, highlighting the consistency of the DLR score with the categorical invasion classifications.

DLR model predicts PitNET recurrence

ARTICLE IN PRESS

We then evaluated the ability of the DLR model to predict tumor recurrence in PitNETs. Using the training cohort median DLR score as a prespecified cutoff, Kaplan–Meier (KM) analyses demonstrated consistent and statistically significant stratification of progression-free survival (PFS) between the low- and high-DLR groups across all cohorts ($P < 0.05$) (**Fig. 3a**).

Next, we further performed subgroup KM and Cox regression analysis adjusting for features to ensure the robustness and reliability of DLR model. Of note, DLR score is a great predictor of tumor recurrence within almost different subgroups of patients stratified by various risk factors **(Supplementary Figures 3-6)**. Cox regression analysis also identified DLR score as an independent prognostic factor for predicting tumor recurrence compared with other risk factors **(Fig. 3b)**.

DLR model indicates aggressive pathology markers

Pathology markers are essential components in determining aggressive behaviors. The DLR score showed a strong association with Ki67 index in the training, independent test cohort, external test cohort and prospective test cohort based on pathological reports (**Fig. 4a**). In the prospective test cohort, IHC results similarly demonstrated that the DLR score differed significantly between tumors with low and high Ki-67 index. (**Fig. 4b**). To eliminate the potential confounding effects of Knosp and Hardy classifications, we

ARTICLE IN PRESS

explored the relationship between the DLR score and Ki-67 index within their respective subgroups. The findings were consistent with the overall results, indicating that the association between the DLR score and Ki-67 index is stable **(Supplementary Figure 7)**. Meanwhile, we further explored the relationship between another pathological marker, p53, and the DLR score. The results similarly showed that tumors with a high DLR score tended to have a higher p53 index (**Fig. 4c**).

Recent studies have indicated that low levels of macrophages within tumors are associated with aggressiveness[25]. Consistently, in both the prospective test cohort and the radiogenomic dataset, we also found that the DLR score was significantly negatively correlated with macrophage biomarkers and infiltration levels at the transcriptomic, proteomic, and cellular levels derived from transcriptomic data using the ImmuCellAI algorithm[26], suggesting that higher DLR scores reflect lower macrophage levels in the tumor, which may be associated with a greater likelihood of aggressive behaviors **(Fig. 4d and Supplementary Figure 8)**.

DLR model reveals aggressive biological processes

The relationship between the DLR score and aggressive BPs was evaluated in the radiogenomics dataset. Differential expression analysis was performed separately within each cluster, comparing samples with high and low DLR scores (**Fig. 5a**). Based on the

ARTICLE IN PRESS

differentially expressed genes, GO analysis revealed enrichment of MAPK signaling pathway, Calcium signaling pathway, TGF-beta signaling pathway, lipid metabolism-related processes and so on in high DLR score group **(Fig. 5b)**. Dysregulated lipid metabolism has been linked to aggressive behaviors in a previous study[27]. In addition, cell adhesion, inflammatory response, etc., were enriched in low DLR score group (**Fig. 5c**). Furthermore, we explored the relationship between the 13 DLFs (f1-f13) and the BPs. The gene set variation analysis (GSVA) and GO analysis categorized the 13 DLFs into four types related to BPs: proliferation (f1, f4, f8), cellular process (f2, f3, f7), immune (f5, f6, f10, f13), and metabolism (f9, f11, f12) (**Fig. 5d**, **Supplementary Data 2**). Overall, the DLR model and relevant DLFs exhibited significant correlations with biological characteristics, providing indications for pathological markers, immune cell, and aggressive BPs. However, invasion classifications (Knosp, Hardy) poorly indicate biological traits, showing weak links to Ki67 index, macrophages, or aggressive BPs, aside from p53 expression. This highlights one of the advantages of the DLR model in evaluating aggressive behaviors over Knosp and Hardy classifications **(Supplementary Figure 9)**.

ARTICLE IN PRESS

The deep learning approach is integrated into an online platform

To facilitate the identification and evaluation of aggressive behaviors, we have incorporated our developed approach into an online platform that offers end-to-end services **(https://mridlr.com/)**. The training cohort, independent test cohort, external test

ARTICLE IN PRESS

cohort and prospective test cohort have been combined into a reference database. MRI images to be analyzed are uploaded to the platform, where backend processing generates the DLR score. These are then compared with the databases to obtain evaluation outcomes **(Fig. 6)**.

Discussion

In this study, we applied a weakly supervised learning strategy based on invasion-related labels, integrated with deep and transfer learning, to construct a model for evaluating aggressive behavior in 1,089 PitNETs. The model achieved robust performance in correlating invasiveness, predicting recurrence, indicating aggressive pathological markers, and revealing aggressive BPs **(Table 2)**. To facilitate clinical use, we developed an end-to-end web platform that streamlined the workflow and enabled preoperative aggressiveness scoring. Patients with high DLR scores, reflecting higher recurrence risk and proliferative activity, may be referred to neurosurgical centers with skull base expertise, while clinicians can consider extended resection strategies to maximize tumor removal. Collectively, this work provides practical tools and evidence for the clinical decision-making and management of aggressive PitNETs.

Aggressiveness is a comprehensive concept involving radiological features, treatment response, prognosis, and biological characteristics. Its assessment is complex and

ARTICLE IN PRESS

challenging[17], as no gold standard or widely accepted clinical biomarker exists. For example, Li et al. identified SF3B1 mutations via genomic sequencing, linking them to therapy resistance and recurrence, and defined them as aggressiveness markers[28]. Similarly, Yan et al. discovered a novel epithelial subpopulation with aggressive signatures using single-cell sequencing; PBK was associated with recurrence and defined as a biomarker of aggressive behaviors[29]. A detailed summary of related studies is in the **Supplementary Table 3**. Each study proposes its own biomarker, making standardization difficult. Most focus on limited aspects, such as invasion and recurrence, and rely on costly approaches unsuitable for widespread or preoperative use.

Therefore, we propose a novel approach integrating MRI with deep learning radiomics to identify aggressive PitNETs and evaluate as many aggressive behaviors as possible based on large cohorts. Following the classic radiomics framework[30,31], the first step is to define aggressive PitNETs using a gold standard, then develop a supervised model mapping MRI features to these labels. However, no reliable or widely validated gold standard or biomarker exists for aggressiveness. Therefore, we integrated weakly supervised invasive labels based on Knosp and Hardy classifications with MRI deep learning to derive a DLR score for evaluating aggressive behaviors. Our findings demonstrate that the DLR model not only correlates with invasion classifications but also effectively predicts prognosis and accurately indicates pathological markers (Ki-67, p53, macrophage) and aggressive

ARTICLE IN PRESS

biological processes. This suggests that the DLR model goes beyond invasiveness and can effectively evaluate PitNET aggressiveness, representing a transition from anatomical invasion assessment to biological aggressiveness evaluation.

The DLR model outperforms invasion classifications in evaluating PitNET aggressiveness for two key reasons. First, invasion (Knosp and Hardy) classifications rely on visual assessment of tumor relations to structures like the internal carotid arteries and optic chiasm, whereas the DLR model integrates these along with tumor size, shape, and MRI signal intensities. Second, its deep learning framework extracts high-dimensional data and captures more comprehensive information beyond the classifications, similar to the emergent abilities seen in large language models.

Our study has several limitations. Although our study assessed many aspects of the aggressive behaviors of PitNETs, it did not fully encompass all relevant features, such as response to medical therapy and tumor growth rate. Second, in predicting tumor recurrence, the DLR score demonstrated robust performance and independent predictive value. However, in certain subgroup analysis for recurrence prediction, a lack of statistically significant differences was observed. A larger dataset needs to be collected, and potential confounding factors require further investigation. Third, because available tumor tissue specimens were required in the prospective test cohort, patients were excluded if specimens were not enough for IHC analysis. This exclusion might potentially bias the study results,

ARTICLE IN PRESS

although no significant difference was observed in many characteristics including age, gender, Knosp grade, and Hardy grade among these cohorts.

In conclusion, we have developed a noninvasive and preoperative deep learning approach for the identification and evaluation of aggressive behaviors in PitNETs based on MRI images. This approach demonstrates excellent performance in evaluating aggressive behaviors of correlating with invasion, predicting recurrence, and indicating biological characteristics, which provides an effective tool for therapeutic decision-making in aggressive PitNETs.

Methods

Data acquisition and study design

This study was approved by the Institutional Review Boards (IRB) of Beijing Tiantan Hospital, Capital Medical University; the First Affiliated Hospital of Zhengzhou University; and Beijing Luhe Hospital, Capital Medical University (approval numbers: KY2020-148-02, 2019-KY-176, and KY2022-155-01). The study was conducted in accordance with the World Medical Association Declaration of Helsinki. For the retrospective cohort, the requirement for informed consent was waived by the IRBs because the data were fully de-identified prior to analysis. Written informed consent was obtained from participants who contributed transcriptomic data and from all individuals

ARTICLE IN PRESS

enrolled in the prospective cohort. Reporting follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

Patients diagnosed with PitNET who underwent preoperative MRI between January 2012 and December 2021 were retrospectively and prospectively enrolled. Four cohorts were collected: training (n=663), independent test (n=172), and prospective test (n=49) from medical center 1 (Beijing Tiantan Hospital, Beijing, China); and external test (n=175 from medical center 2, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China; n=20 from medical center 3, Beijing Luhe Hospital, Beijing, China) **(Table 1)**. The training cohort included 106 patients with transcriptomic data, which formed a radiogenomic dataset for exploring the association between the model and underlying biological processes. The detailed collection process is shown in **Supplementary Figure 10.** Data included MRI data, clinicopathologic data (age, gender, symptoms, transcription factors, hormones, Ki-67), transcriptomics, tumor specimens, postoperative extent of resection, and follow-up. Transcriptomic data for the radiogenomic dataset were obtained via second-generation sequencing. Paraffin-embedded tissue was used for IHC in the prospective test cohort. The model was built using Python 3.8.5 and PyTorch 1.10.2 on a server with Platinum82 (80 vCPUs, 2.5 GHz) and two NVIDIA Tesla V100 SXM2 GPUs (CentOS 7.8).

ARTICLE IN PRESS

MRI images acquisition and processing

All MRI were acquired on 1.5T and 3.0T MRI scanners with 24-channel head matrix coils (Magnetom Skyra or Trio TIM, Siemens Healthcare, or Discovery MR 750, GE). The brain imaging protocol included the following sequences: (1) T1-weighted imaging (T1W); (2) T2-weighted imaging (T2W); (3) contrast-enhanced T1-weighted imaging (T1C). T1C sequences were acquired immediately after intravenous administration of gadolinium-based contrast agent at a dose of a 0.1 mmol per kilogram with the same parameters as the precontrast T1 sequence. Prior foundational studies have demonstrated that rigorous preprocessing of MRI data enhances image quality and improves downstream task performance[32,33]. With reference to prior work, we implemented a standardized preprocessing pipeline to reduce inter-patient intensity and geometric variability. First, the non-uniformity in low-frequency intensity was adjusted via the N4 bias correction algorithm, as implemented in the Advanced Normalization Tools (ANTs). Second, the image data were subjected to reorientation to the left posterior-superior (LPS) coordinate system. Third, all voxels were resampled to into $1\times1\times1mm^3$ using bicubic interpolation. Next, rigid registration was performed to align the three MR sequences using axial resampled T1C as a template.

Assignment of Knosp and Hardy grades on preoperative MRI

ARTICLE IN PRESS

The Knosp and Hardy labels were assigned based on preoperative MRI. Two experienced readers, blinded to surgery, pathology, and outcomes, independently graded all pre-operative MRIs. Coronal T1C images were the primary sequence for cavernous-sinus assessment, with axial/sagittal T1C and T2 as adjuncts. Knosp grades (0–4) were assigned using the intracavernous ICA tangent-line method. Hardy/Hardy–Wilson grading included sellar floor destruction (0–IV) and extrasellar extension (A–E)[6,34,35]. Disagreements were resolved by consensus with a senior reviewer. Given the small number of Knosp 0 grade PitNET samples (16 cases in the training cohort, 3 cases in the radiogenomics dataset, 5 cases in the independent test cohort, 10 cases in the external test cohort, and 0 cases in the prospective test cohort), this subgroup was combined with Knosp 1 grade samples for further analysis.

Assessment of progression-free survival and postoperative residual tumor

PFS was defined from date of the first surgical resection to the date of tumor progression, or/and hormonal recurrence during the follow-up. In that, tumor progression was defined as significant enlargement (>2 mm in any direction) of the tumor remnants or the appearance of new masses detected by MRI during the follow-up. Hormonal recurrence based on lab testing was defined the level of serum hormonal markedly increases again when PitNETs reached a state of hormonal decrease after the first surgical resection. There are three detailed conditions as follows: (1). significant enlargement (>2 mm in any

ARTICLE IN PRESS

direction) of the residual tumor after incomplete resection (**Supplementary Figure 11a**); (2). the appearance of new masses after complete resection with no residual tumor on MRI during the follow-up (**Supplementary Figure 11b**); (3). hormonal recurrence even if the residual tumors are stable (**Supplementary Figure 11c**).

Differentiation of postoperative residual PitNETs primarily involves distinguishing residual tumor from postoperative hematoma. There are three main methods to differentiate hematoma from residual tumor. Hematoma on the first postoperative MRI (day 3) appears hyperintense on T1W, hypointense on T2W, shows no enhancement on contrast-enhanced T1W, and has an irregular shape. Residual PitNET tumor shows MRI signals like the preoperative tumor and often represents part of the original tumor on preoperative MRI. If differentiation remains uncertain, we reassess with contrast-enhanced MRI at 3 months postoperatively.

ARTICLE IN PRESS

RNA sequencing and immunohistochemistry

Total RNA was extracted from surgical fresh-frozen specimens using RNAiso Plus (TaKaRa, Dalian, China). RNA quantity and integrity were assessed, index-coded libraries were constructed, and samples were sequenced on an Illumina HiSeq 3000 platform to generate 150 bp paired-end reads. Read quality was checked with FastQC, clean reads were

ARTICLE IN PRESS

aligned to the reference genome using STAR, and gene expression was quantified with HTSeq (v0.6.0) as fragments per kilobase of transcript per million mapped reads (FPKM).

For immunohistochemistry, fresh tissue was fixed for >24 h, processed through graded alcohols, embedded in paraffin, and sectioned at 4 μm. After deparaffinization and rehydration, sections were incubated overnight at 4°C with primary antibodies against Ki67 (ZSGB-BIO, ZM-0166), p53 (ZSGB-BIO, ZM-0408), and CD68 (ZSGB-BIO, ZM-0060), followed by the appropriate secondary antibody (ZSGB-BIO, PV-9002) for 1 h at room temperature. Staining was visualized with DAB and counterstained with hematoxylin. Quantitative evaluation was performed on the whole-section area, and the percentages of Ki67- and CD68-positive cells and the immunoreactive score (IRS) of p53 were calculated.

Tumor segmentation

Tumor masks on T1C images were delineated by an experienced neuro-radiologist using ITK-SNAP and manually verified by another radiologist blinded to patient data. Discrepancies were resolved by consensus. A total of 234 PitNET masks formed the training dataset. Three 3D deep learning models—nnUnet, nnFormer, and Unet—were trained on this dataset using the Dice coefficient for optimization, with 5-fold cross-validation applied. The optimal model, based on Dice scores, was used for automatic segmentation of remaining cases **(Supplementary Figure 2)**. Resulting masks were

ARTICLE IN PRESS

reviewed and adjusted by two radiologists, then applied to T1W and T2W images after careful evaluation and modifications.

Fine-tuned deep network for feature extraction

First, the data were adapted to match the input format of the deep network by converting 3D MRI images into three-channel 2D slices for analysis, selecting three axial cross-sections for each modality (T1W, T2W, and T1C), consisting of the center slice containing the carotid artery and the adjacent ± 2 slices **(Fig. 1c)**. Second, the slices were cropped, resized to 224×224 pixels, normalized, and reconstituted as three-channel RGB-like images, thereby constructing the ROIs for each patient, namely ROI1, ROI2, and ROI3. Next, the three ROIs (ROI1, ROI2, ROI3) from each patient were treated as individual samples, each inheriting the patient-level weak label, and were then input into the deep learning classification models. Third, these ROIs were input into ResNet50, Vision Transformer, and Swin Transformer models pretrained on ImageNet, with parameters fine-tuned using cross-entropy loss.

The weakly supervised label was invasion type based on Knosp and Hardy classification. To ensure that the weakly supervised invasion labels more closely capture aggressiveness, we adopted stricter invasion criteria compared with previous studies[6,36]. Specifically, PitNETs classified as Knosp grade I and Hardy grades I, II, and/or stages A, B, and C were

ARTICLE IN PRESS

categorized as non-invasive, whereas those classified as Knosp grade IV and Hardy grades III, IV, and/or stages D and E were deemed invasive. Thus, a total of 346 PitNETs with other grades such as Knosp 2 or 3 were not used as training samples. Finally, a total of 118 were invasive PitNETs as positive label, and 199 were not invasive samples as negative label in the training cohort. 85% of labeled cases formed the training dataset, and 15% were reserved for testing. The Adam optimizer was used, with hyperparameters selected based on overfitting and underfitting assessments. The best-performing model was used to extract DLFs.

As a result, three deep learning frameworks (Resnet50, Vision transformer, and Swin transformer) were employed for the classification of PitNET invasion, among which the transfer-learning Swin transformer (TLST) achieved the best performance. We then tested the TLST of hyperparameters including type, learning rate, and batch size, revealing that the best sets were small, 0.0005, and 128, respectively (**Supplementary Figure 1, Supplementary Figure 2f, Supplementary Tables 1-2**). The confusion matrices showed that PitNETs were highly accurately classified (**Supplementary Figure 2g**). Thus, TLST with the above hyperparameters was selected as the tumor classification model to extract DLFs.

DLFs selection and DLR signature building

ARTICLE IN PRESS

Considering the downstream analysis, the processed three ROIs of each patient in this study were fed into the tumor classification model, and the DLFs of the penultimate layer were extracted. Z-score method was employed to normalize all the features into a standardized value range within each cohort separately.

Feature selection methods including *t*-test, mutual information, recursive feature elimination, and LASSO were employed to identify DLFs correlated with tumor aggression. First, *t*-test and mutual information (MI) method were employed for feature selection. Only features with MI value greater than the pre-defined threshold values were chosen. According to the pre-experiment, the threshold values were considered as hyperparameters and set at 0.1, 0.15, 0.2, 0.25. Next, because the number of deep learning features was relatively large when compared with the number of patients, we applied recursive feature elimination (RFE) and least absolute shrinkage and selection operator (LASSO) to select the most significant and representative features. A RFE algorithm was used for feature selection. The RFE algorithm method attempts to find the best subset of size σ through a greedy backward selection. Similarly, σ was also considered a hyperparameter and set at 10, 15, 20, 25. LASSO involves penalizing irrelevant variables to zero and only retains useful features, therefore effectively reducing the number of variables. To provide robust generalized performance of a model that best fits the observed data, five-fold cross-validation was performed using default parameters. Next, multivariable logistic regression

ARTICLE IN PRESS

analysis was conducted to build the deep learning radiomic model (DLR model) based on the screened features by the methods above. In this study, except for MI value and σ, the other parameters were chosen by default. The hyperparameter selection of the two values above was determined by the discrimination between positive and negative label of the multivariable logistic regression analysis. In the end, the hyperparameters for the MI value and σ were set at 0.2 and 20, respectively, and were optimized for the best performance.

Overall, a multivariable logistic regression analysis was conducted to build a DLR model and calculate DLR scores for each patient. The DLR scores were ranked, and cases were dichotomized into high and low groups using the median DLR score as the cutoff for subsequent analyses. The cut-off values from the training cohort were then applied to other cohorts.

ARTICLE IN PRESS

Radiogenomics studies

In the radiogenomic dataset, patients from the radiogenomic dataset were divided into subgroups by unsupervised clustering for further differential analysis. Consensus clustering was performed to identify transcriptome clusters of PitNETs using the ConsensusClusterPlus package. The top 1500 most variable genes were selected, based on their standard deviation. Parameters were reps=1000, pFeature = 0.8, pItem = 0.8, clusterAlg = “hc”, distance = “spearman” in the range of 2 to 10 clusters. The consensus

ARTICLE IN PRESS

matrix of k = 3 clusters was shown in our study **(Supplementary Figure 12a)**. The delta plot of the relative change in the area under the cumulative distribution function (CDF) curve was calculated to estimate the average pairwise consensus matrix within consensus clusters **(Supplementary Figure 12b-12c).** We then determined the consensus matrix of k = 3 as the best solution for clustering as it presents the most separated clusters compared with other k values. Through unsupervised consensus clustering, the radiogenomics dataset was divided into three cell lineages clusters: SF1, TPIT, and PIT1 (**Supplementary Figure 12)**. The composition of three transcription factor (TF) clusters is detailed in **Supplementary Table 4.**

PitNETs typically show cytodifferentiation toward one of three pituitary TF-defined lineages: SF1, TPIT or PIT1. To minimize confounding from lineage differences[16], differential analysis were performed within each TF subgroup, comparing high vs. low DLR scores, Knosp grades 3-4 vs. 1-2, and Hardy invasion vs. non-invasion. Enrichment analysis was conducted to further clarify biological processes related to the intersection of all differential genes using Gene Ontology (GO) and KEGG tools via the Metascape database and R package "clusterProfiler" ($q < 0.05$). Finally, GSVA and GO were used to annotate the biological relevance of DLR features, identifying significant pathway associations ($P < 0.05$) and confirming enrichment reliability ($q < 0.05$).

Statistical analysis

ARTICLE IN PRESS

Statistical analysis was performed using R 3.5.0 and Python 3.8.5. Categorical and continuous variables were analyzed with Fisher's exact test, Chi-square test, *t*-test, or Mann-Whitney U test. Spearman and Kendall correlation evaluated variable concordance. Kaplan-Meier survival analysis employed log-rank tests, and $P < 0.05$ indicated significance.

Data availability

All data in the article and supplementary information can be provided upon reasonable request to the corresponding author.

Code availability

All code used for modeling and data analysis is deposited in a publicly accessible repository (https://github.com/ttyywyy/identifying-and-evaluating-aggressive-behaviors-in-PitNETs).

Acknowledgements

This project was supported by the National Natural Science Foundation of China (No. 82071996, No. 81802483, No. 62106266 and No. 82003075), China Postdoctoral Science Foundation (No. 2024M762180), and Beijing Postdoctoral Research Foundation.

ARTICLE IN PRESS

Author contributions

W.J. and C.Z. conceived the clinical idea for this project. C.Z. and Y.T. designed the deep learning algorithm architecture. Y.W. wrote the code and analyzed the data. Y.W. and C.Z. blindly assessed invasion, and W.J. was the third neuro-radiologists for consensus grading. X.G., S.M., J.Y., Y.Z., L.M., D.L. and W.J. conducted all the operations. C.Z., X.G., Y.L., Y.S. and L.M. assisted with manuscript preparation, data/image analysis, and manuscript revision. D.L., X.G. and Y.S. took part in the discussion. All authors participated in the interpretation of the reported experiments or results. All the authors reviewed the final version of the manuscript and were accepted as a co-author.

Competing interests

The authors declare no competing interests.

ARTICLE IN PRESS

ARTICLE IN PRESS

References

1. Fernandez A, Karavitaki N & Wass JA. Prevalence of pituitary adenomas: a community-based, cross-sectional study in Banbury (Oxfordshire, UK). *Clin Endocrinol (Oxf)* **72**, 377-382 (2010).
2. Melmed S et al. Diagnosis and treatment of hyperprolactinemia: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab* **96**, 273-288 (2011).
3. Villa C et al. A standardised diagnostic approach to pituitary neuroendocrine tumours (PitNETs): a European Pituitary Pathology Group (EPPG) proposal. *Virchows Arch* **475**, 687-692 (2019).
4. Raverot G et al. Aggressive pituitary tumours and pituitary carcinomas. *Nat Rev Endocrinol* **17**, 671-684 (2021).
5. Mohr G, Hardy J, Comtois R & Beauregard H. Surgical management of giant pituitary adenomas. *Can J Neurol Sci* **17**, 62-66 (1990).
6. Knosp E, Steiner E, Kitz K & Matula C. Pituitary adenomas with invasion of the cavernous sinus space: a magnetic resonance imaging classification compared with surgical findings. *Neurosurgery* **33**, 610-617; discussion 617-618 (1993).
7. Micko AS, Wöhrer A, Wolfsberger S & Knosp E. Invasion of the cavernous sinus space in pituitary adenomas: endoscopic verification and its correlation with an MRI-based classification. *J Neurosurg* **122**, 803-811 (2015).
8. CB W. Neurosurgical Management of Large and Invasive Pituitary Tumors. In: GT Tindall, WF Collins (eds). Clinical Management of Pituitary Disorders. *New York: Raven Press*, 335–342 (1979).
9. Ghadir M et al. Cell proliferation, apoptosis, and angiogenesis in non-functional pituitary adenoma: association with tumor invasiveness. *Endocrine* **69**, 596-603 (2020).
10. Grimm F et al. Ki-67 labeling index and expression of p53 are non-predictive for invasiveness and tumor size in functional and nonfunctional pituitary adenomas. *Acta Neurochir (Wien)* **161**, 1149-1156 (2019).
11. Sarkar S, Chacko AG & Chacko G. An analysis of granulation patterns, MIB-1 proliferation indices and p53 expression in 101 patients with acromegaly. *Acta Neurochir (Wien)* **156**, 2221-2230; discussion 2230 (2014).
12. Raverot G et al. European Society of Endocrinology Clinical Practice Guidelines for the management of aggressive pituitary tumours and carcinomas. *Eur J Endocrinol* **178**, G1-g24 (2018).
13. Mete O & Lopes MB. Overview of the 2017 WHO Classification of Pituitary Tumors. *Endocr Pathol* **28**, 228-243 (2017).

ARTICLE IN PRESS

14. McCormack A et al. Treatment of aggressive pituitary tumours and carcinomas: results of a European Society of Endocrinology (ESE) survey 2016. *Eur J Endocrinol* **178**, 265-276 (2018).
15. Kemeny HR et al. Targeting PD-L1 Initiates Effective Antitumor Immunity in a Murine Model of Cushing Disease. *Clin Cancer Res* **26**, 1141-1151 (2020).
16. Neou M et al. Pangenomic Classification of Pituitary Neuroendocrine Tumors. *Cancer Cell* **37**, 123-134.e125 (2020).
17. Asa SL et al. From pituitary adenoma to pituitary neuroendocrine tumor (PitNET): an International Pituitary Pathology Club proposal. *Endocr Relat Cancer* **24**, C5-c8 (2017).
18. Feng CM et al. DONet: Dual-Octave Network for Fast MR Image Reconstruction. *IEEE Trans Neural Netw Learn Syst* **36**, 3965-3975 (2025).
19. Feng CM et al. Multimodal Transformer for Accelerated MR Imaging. *IEEE Trans Med Imaging* **42**, 2804-2816 (2023).
20. Wang Y et al. Factors Influencing Hormone Remission in Growth Hormone-Secreting Pituitary Neuroendocrine Tumors With Residual Tumor: A Retrospective Cohort Study. *CNS Neurosci Ther* **31**, e70574 (2025).
21. Park YW et al. Radiomics With Ensemble Machine Learning Predicts Dopamine Agonist Response in Patients With Prolactinoma. *J Clin Endocrinol Metab* **106**, e3069-e3077 (2021).
22. Zhang LL et al. Pretreatment MRI radiomics analysis allows for reliable prediction of local recurrence in non-metastatic T4 nasopharyngeal carcinoma. *EBioMedicine* **42**, 270-280 (2019).
23. Ugga L et al. Prediction of high proliferative index in pituitary macroadenomas using MRI-based radiomics and machine learning. *Neuroradiology* **61**, 1365-1373 (2019).
24. Peng A et al. A machine learning model to precisely immunohistochemically classify pituitary adenoma subtypes with radiomics based on preoperative magnetic resonance imaging. *Eur J Radiol* **125**, 108892 (2020).
25. Lin S et al. Single-cell transcriptomics reveal distinct immune-infiltrating phenotypes and macrophage-tumor interaction axes among different lineages of pituitary neuroendocrine tumors. *Genome Med* **16**, 60 (2024).
26. Miao YR et al. ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Adv Sci (Weinh)* **7**, 1902880 (2020).
27. Zhang Q et al. Single-cell sequencing identifies differentiation-related markers for molecular classification and recurrence prediction of PitNET. *Cell Rep Med* **4**, 100934 (2023).

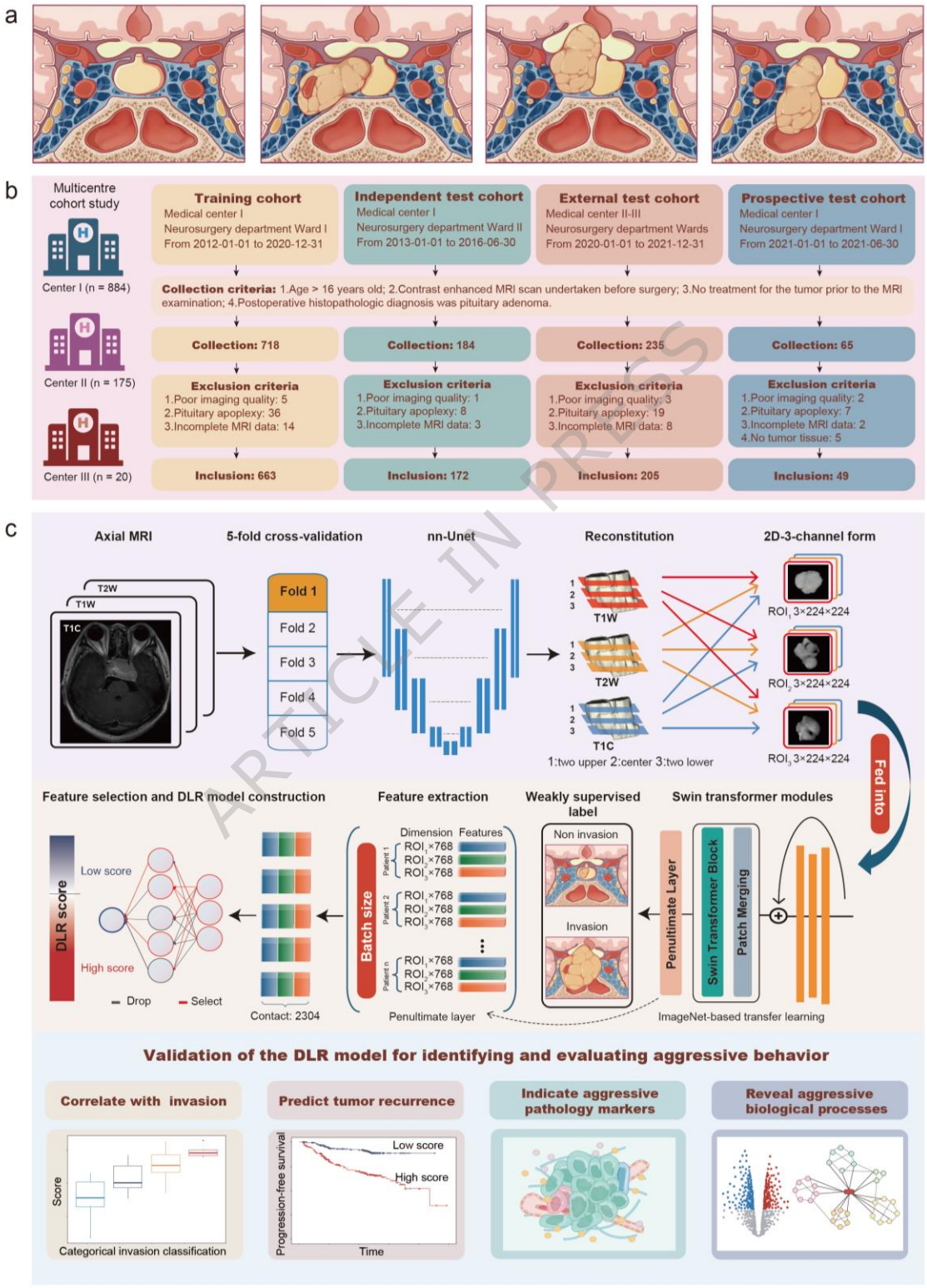
ARTICLE IN PRESS

28. Li C et al. Somatic SF3B1 hotspot mutation in prolactinomas. *Nat Commun* **11**, 2506 (2020).
29. Yan N et al. Single-cell transcriptomic analysis reveals tumor cell heterogeneity and immune microenvironment features of pituitary neuroendocrine tumors. *Genome Med* **16**, 2 (2024).
30. Jiang Y et al. Non-invasive tumor microenvironment evaluation and treatment response prediction in gastric cancer using deep learning radiomics. *Cell Rep Med* **4**, 101146 (2023).
31. Ma C et al. Multimodal-based machine learning strategy for accurate and non-invasive prediction of intramedullary glioma grade and mutation status of molecular markers: a retrospective study. *BMC Med* **21**, 198 (2023).
32. Feng C-M et al. Task transformer network for joint MRI reconstruction and super-resolution. In: *International conference on medical image computing and computer-assisted intervention*). Springer (2021).
33. Feng C-M, Fu H, Yuan S & Xu Y. Multi-contrast MRI super-resolution via a multi-stage integration network. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*). Springer (2021).
34. Hardy J & Vezina JL. Transsphenoidal neurosurgery of intracranial neoplasm. *Adv Neurol* **15**, 261-273 (1976).
35. Melmed S et al. Clinical Biology of the Pituitary Adenoma. *Endocr Rev* **43**, 1003-1037 (2022).
36. Araujo-Castro M et al. Radiological Knosp, Revised-Knosp, and Hardy-Wilson Classifications for the Prediction of Surgical Outcomes in the Endoscopic Endonasal Surgery of Pituitary Adenomas: Study of 228 Cases. *Front Oncol* **11**, 807040 (2021).

ARTICLE IN PRESS

Figures

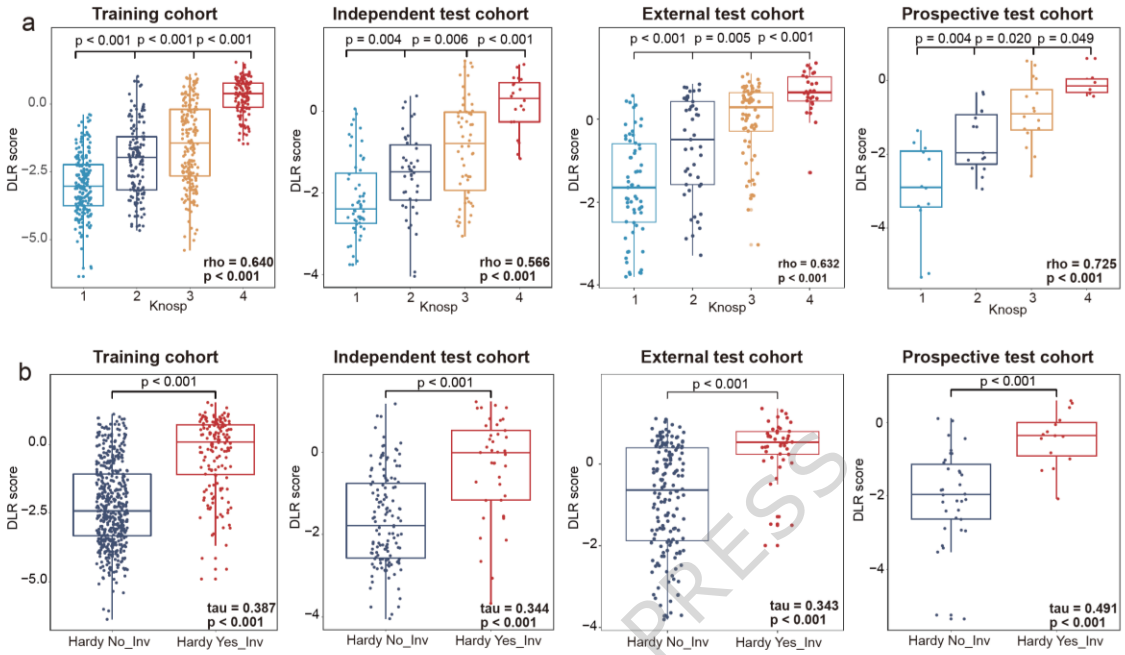
Fig.1



ARTICLE IN PRESS

ARTICLE IN PRESS

514 **Fig.2**



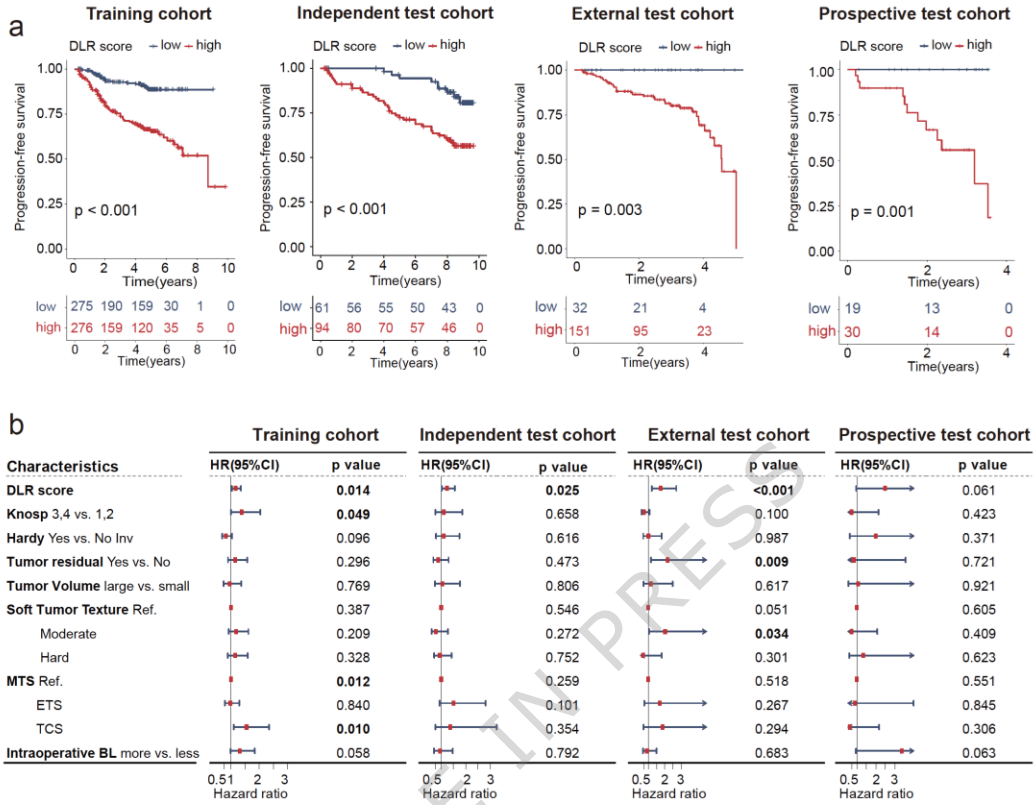
ARTICLE IN PRESS

515

516

ARTICLE IN PRESS

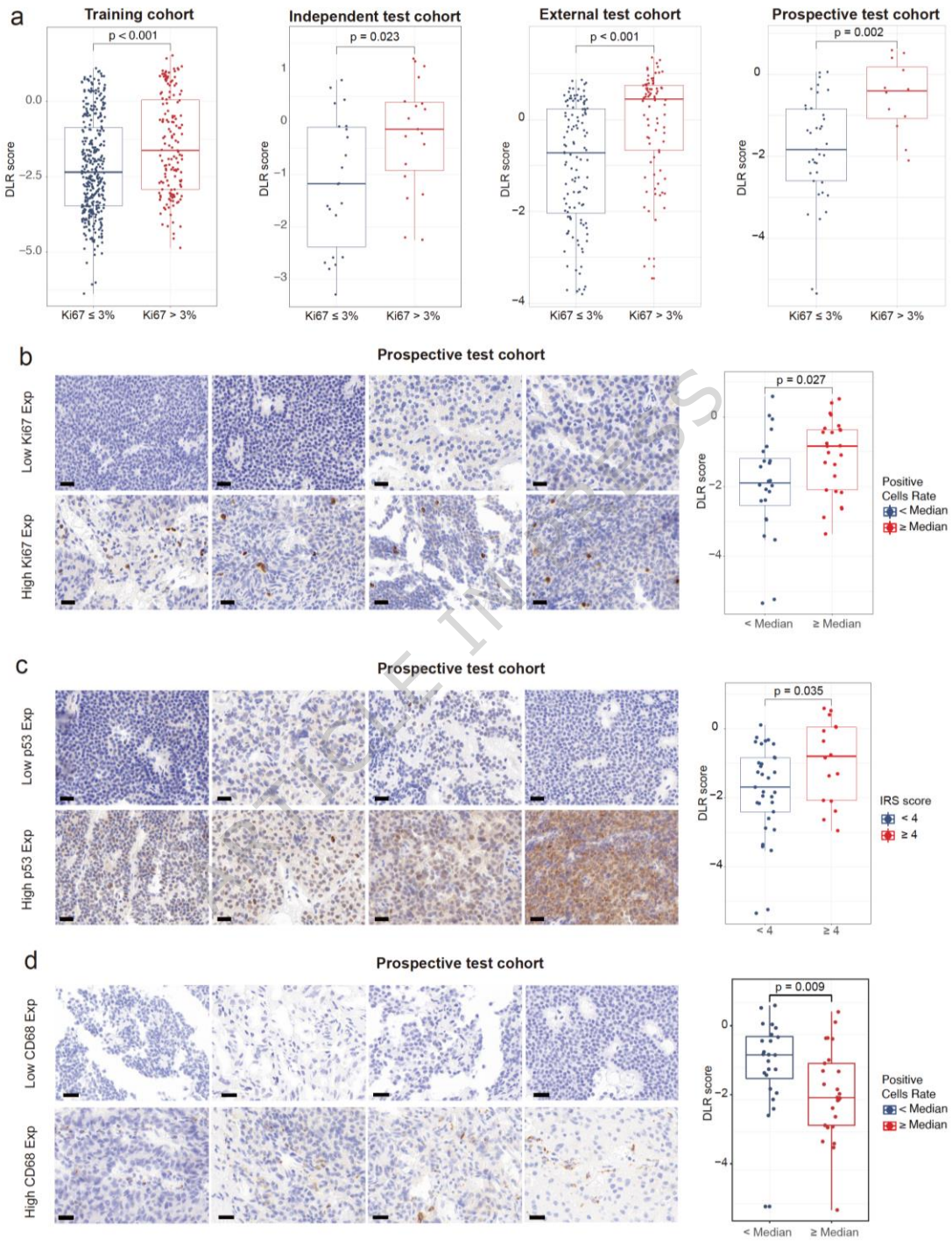
517 **Fig.3**



518

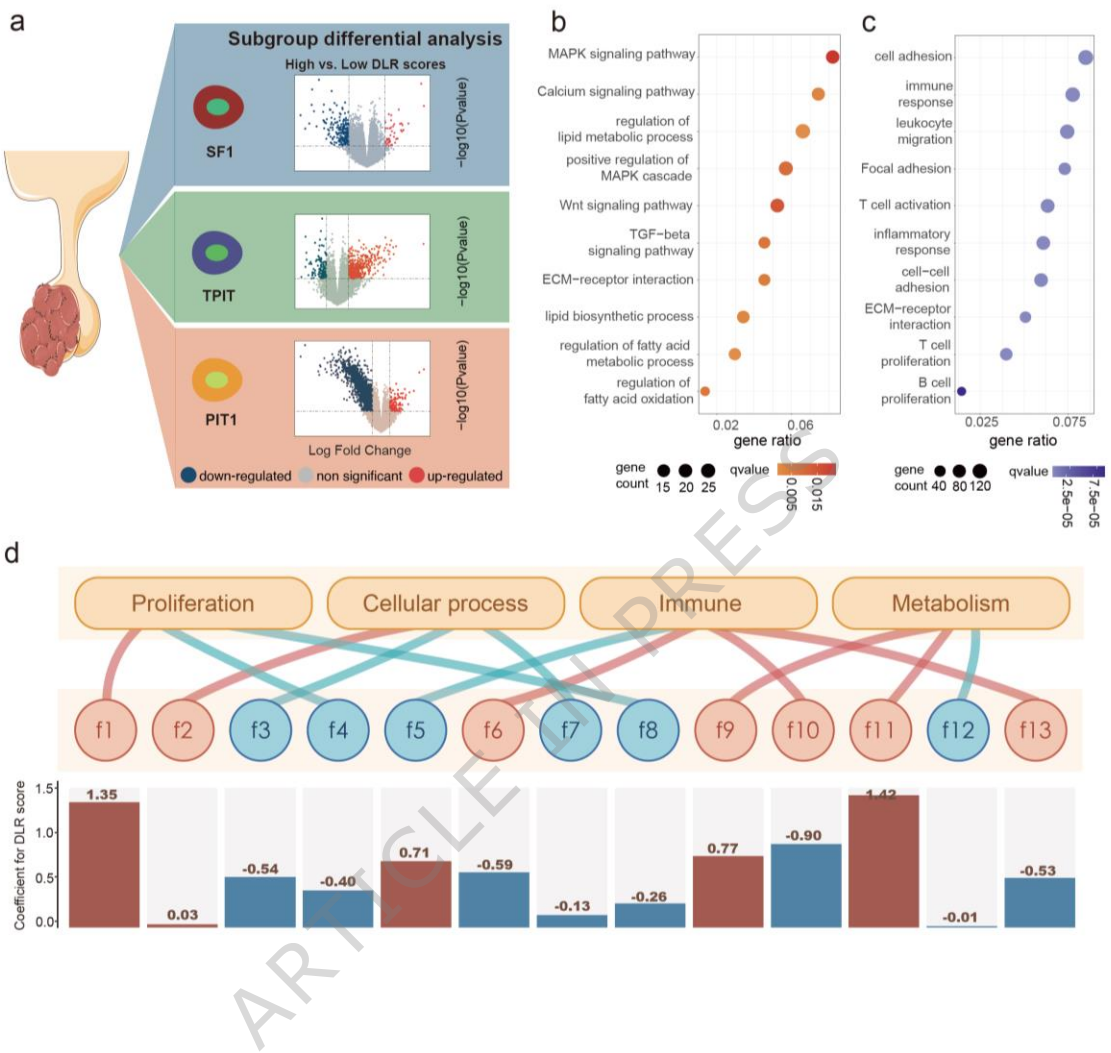
519

520 **Fig.4**


a
Training cohort
p < 0.001
Independent test cohort
p = 0.023
External test cohort
p < 0.001
Prospective test cohort
p = 0.002
DLR score
Ki67 ≤ 3%
Ki67 > 3%
b
Prospective test cohort
Low Ki67 Exp
High Ki67 Exp
p = 0.027
< Median
≥ Median
Positive Cells Rate
c
Prospective test cohort
Low p53 Exp
High p53 Exp
p = 0.035
< 4
≥ 4
IRS score
d
Prospective test cohort
Low CD68 Exp
High CD68 Exp
p = 0.009

521

522 **Fig.5**

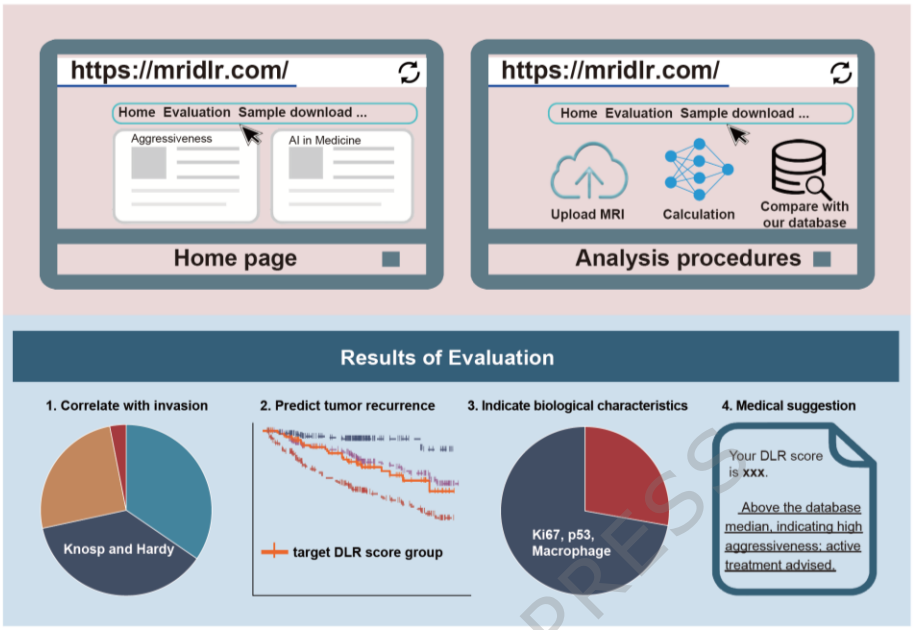


ARTICLE IN PRESS

523

ARTICLE IN PRESS

524 **Fig.6**



ARTICLE IN PRESS

525

ARTICLE IN PRESS

526 **Tables**

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 1. Baseline Characteristics of the Cohorts

Characteristics	Training cohort (n = 663)	Independent test cohort (n = 172)	External test cohort (n = 205)	Prospective test cohort (n = 49)	P value
Age (yrs)	47.9±12.22	47.65±11.98	47.33±12.85	47.80±11.04	0.949
Gender					
female	324 (48.9%)	77 (44.8%)	106 (51.7%)	28 (57.1%)	0.361
male	339 (51.1%)	95 (55.2%)	99 (48.3%)	21 (42.9%)	
Initial symptoms					
headache	184 (27.8%)	51 (29.7%)	65 (31.7%)	8 (16.3%)	0.004
acromegaly	71 (10.7%)	10 (5.8%)	14 (6.8%)	2 (4.1%)	
vision deficits	237 (35.7%)	63 (36.6%)	56 (27.4%)	19 (38.8%)	
others	171 (25.8%)	48 (27.9%)	70 (34.1%)	20 (40.8%)	
Knosp					
1	205 (30.9%)	56 (32.6%)	63 (30.7%)	12 (24.5%)	0.443
2	141 (21.3%)	42 (24.4%)	45 (21.9%)	15 (30.6%)	
3	199 (30.0%)	54 (31.4%)	69 (33.7%)	16 (32.7%)	
4	118 (17.8%)	20 (11.6%)	28 (13.7%)	6 (12.2%)	
Hardy-Wilson					
No-Inv	505 (76.2%)	131 (76.2%)	160 (78.0%)	34 (69.4%)	0.925
Yes-Inv	158 (23.8%)	41 (23.8%)	45 (22.0%)	15 (30.6%)	
Surgery approach					
MTS	289 (43.6%)	52 (30.2%)	34 (16.6%)	22 (44.9%)	<0.001
ETS	270 (40.7%)	99 (57.6%)	160 (78.0%)	19 (38.8%)	
TCS	104 (15.7%)	21 (12.2%)	11 (5.4%)	8 (16.3%)	
IHC hormone					
PRL	245 (38.0%)	30 (17.4%)	66 (32.2%)	12 (24.5%)	<0.001
GH	158 (23.8%)	29 (16.9%)	48 (23.4%)	8 (16.3%)	
TSH	84 (12.7%)	14 (8.1%)	20 (9.8%)	4 (8.2%)	
FSH	193 (29.1%)	38 (22.1%)	58 (28.3%)	23 (46.9%)	
LH	37 (5.6%)	9 (5.2%)	32 (15.6%)	36 (73.5%)	
ACTH	159 (24.0 %)	24 (14.0%)	39 (19.0%)	10 (20.4%)	
NON	138 (20.8%)	73 (42.4%)	54 (26.3%)	3 (6.1%)	
Unknown	7 (1.1%)	1 (0.6%)	2 (1.0%)	0 (0.0%)	
IHC TF[#]					
PIT1	93 (14.0%)	5 (2.9%)	58 (28.3%)	17 (34.7%)	<0.001
TPIT	42 (6.3%)	6 (3.5%)	46 (22.4%)	8 (16.3%)	
SF1	82 (12.4%)	13 (7.6%)	44 (21.5%)	34 (69.4%)	
NON	1 (0.2%)	0 (0.0%)	3 (1.5%)	0 (0.0%)	
Unknown	499 (75.3%)	150 (87.2%)	86 (42.0%)	2 (4.1%)	

Invasion (Inv); microscopic transsphenoidal surgery (MTS), endoscopic transsphenoidal surgery (ETS), transcranial surgery (TCS); #: Many unknowns remain because the transcription factor classification was only formally introduced by the WHO in 2017, and large-scale clinical applications have only gradually started since 2018.

ARTICLE IN PRESS

Table 2. Summary of different evaluation contents

Evaluation contents	Developed model	Validation Reference	Cohort
Correlate with invasion	DLR model	Knosp and Hardy classification	Training cohort Independent test cohort External test cohort Prospective test cohort
Predict tumor recurrence	DLR model	Follow-up outcomes including three conditions[#]	Training cohort Independent test cohort External test cohort Prospective test cohort
Indicate aggressive pathology markers	DLR model	Immunohistochemistry- and transcriptomics data	Training cohort Independent test cohort External test cohort Prospective test cohort
Reveal biological characteristics	DLR model	Transcriptomic data	Radiogenomics dataset

[#]: Details of three conditions can be found in **Supplementary Figure 11**.

527

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure legends

Fig. 1 | Overview of our study. (a) The graph shows the normal pituitary gland and PitNETs with invasion to the lateral, superior, and inferior regions. **(b)** Patient collection process of the study. **(c)** Workflow of our study.

Fig. 2 | The performance of the DLR model in evaluating PitNET invasion. (a-b) Boxplots showing DLR score distributions by Knosp and Hardy-Wilson classifications across the four cohorts. No_Inv: non-invasion; Yes_Inv: invasion.

Fig. 3 | The performance of the DLR model in predicting tumor recurrence. (a) Kaplan-Meier analysis of progression-free survival across the four cohorts. **(b)** Multivariate Cox regression analysis of progression-free survival across the four cohorts. MTS: microscopic transsphenoidal surgery; ETS: endoscopic transsphenoidal surgery; TCS: transcranial surgery; BL: blood loss.

Fig. 4 | The performance of the DLR model in indicating pathological markers. (a) Boxplot of DLR score between low- and high Ki67 expression. **(b-d)** IHC results and correlations between DLR scores and Ki-67, p53, or CD68 in the prospective test cohort. Scale bars, 20 μm. No_Inv: non-invasion; Yes_Inv: invasion. Exp: expression.

ARTICLE IN PRESS

Fig. 5 | The performance of the DLR model in revealing aggressive biological processes. (a) Volcano plots showing differential mRNA expression between high- and low-DLR score groups across three transcription factor subgroups (red: upregulated; blue: downregulated). **(b, c)** Enriched pathways for upregulated **(b)** and downregulated **(c)** mRNAs are shown as red/blue dots. **(d)** Correlations between DLR features and biological processes (red: positive; blue: negative) with bar charts showing feature coefficients (red: positive; blue: negative).

Fig. 6 | Integration of deep learning approach into a web platform for convenient clinical application. By uploading MRI data, performing computation, and comparing the results with the database, users can obtain an evaluation of aggressive behaviors and a DLR score, along with corresponding medical suggestions.

ARTICLE IN PRESS