frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 23 September 2025
DOI 10.3389/fonc.2025.1657604

Check for updates

OPEN ACCESS

EDITED BY
Shuai Ren,
Affiliated Hospital of Nanjing University of Chinese Medicine, China

REVIEWED BY
Andrea Bianconi,
University of Genoa, Italy
Prabhishek Singh,
Bennett University, India
Qiang Wen,
Stanford University, United States
Hamidreza Sadeghsalehi,
Imperial College, United Kingdom

*CORRESPONDENCE
Lingchun Xu
1308022992@qq.com
Qinglei Zhang
13915961989@163.com
Na Yin
2008.yinna@163.com

†These authors have contributed equally to this work

RECEIVED 01 July 2025
ACCEPTED 01 September 2025
PUBLISHED 23 September 2025

CITATION
Shi D, Yang M, Dong M, Xuan N, Zhu Y, Lv X, Xie C, Xia F, Xu L, Zhang Q and Yin N (2025) Development and validation of a deep learning model using MR imaging for predicting brain metastases: an accuracy-focused study. *Front. Oncol.* 15:1657604. doi: 10.3389/fonc.2025.1657604

COPYRIGHT
© 2025 Shi, Yang, Dong, Xuan, Zhu, Lv, Xie, Xia, Xu, Zhang and Yin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Development and validation of a deep learning model using MR imaging for predicting brain metastases: an accuracy-focused study

Dan Shi[1†], Meng Yang[2†], Min Dong[3†], Ning Xuan[4], Yinsu Zhu[1], Xiaoqiong Lv[1], Chao Xie[1], Fei Xia[1], Lingchun Xu[1*], Qinglei Zhang[2*] and Na Yin[1*]

[1]Department of Radiology, Jiangsu Cancer Hospital, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing, China, [2]Department of Radiology, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, China, [3]The People's Hospital of DanYang Oncology, Zhenjiang, China, [4]Independent Researcher, Los Angeles, CA, United States

Background: Brain metastases (BM), originating from extracranial malignancies, significantly threaten patient health. Accurate BM identification is crucial but labor-intensive manually. This study developed and validated a system for BM diagnosis, assessing its performance and stability.

Methods: 470 patients diagnosed with BM were divided into an 80% training set (n=379) and a 20% internal test set (n=91) using systematic sampling. An additional 172 patients were retrospectively enrolled for external validation. A comprehensive preprocessing pipeline was implemented. We developed a 3D U-Net model with a ResNet-34 backbone for BM prediction. MRI scans were resampled to 0.833 mm^3 isotropic voxels, underwent skull stripping using SynthStrip, and were intensity-normalized via Z-score normalization. The model was trained on MRI scans paired with segmentation masks, utilizing ImageNet-pretrained encoder weights and a patch-based strategy (128×128×128 voxels).

Results: The model maintained perfect specificity and AUCs across gender and age groups, with no significant differences in other metrics, confirming false positive exclusion unaffected by demographics. By cancer type: Internal testing showed significant difference of AUC (p<0.001) between lung cancer (n=74) and other cancers (n=17). The differences of other performance metrics were not statistically significant (p>0.13), though other cancers showed higher median F1/IoU/MCC. External validation showed other cancers (n=79) had significantly higher precision than lung cancer (n=93) (p<0.05). Lung cancer AUC (0.82) was significantly lower than other cancers (0.89) (p<0.001), suggesting need for sensitivity optimization; both maintained specificity=1.0000. Model time was

significantly shorter than manual annotation (internal: 69s vs 113s; external: 66s vs 96s; both $p<0.001$), with high agreement.
Conclusion: The model demonstrated strong robustness and perfect specificity across demographics. While showing cancer type dependency (requiring improved lung cancer sensitivity), its high efficiency (40%-50% time reduction) and generalization provide a solid foundation for clinical translation.

KEYWORDS
brain metastases, deep learning, artificial intelligence, diagnostic accuracy, magnetic resonance imaging

1 Introduction

Brain metastases (BM) are malignant tumors originating from extracranial primary tumors that metastasize to the brain parenchyma. Representing the most common type of intracranial tumor in adults, BM occur in approximately 10% to 40% of patients with solid tumors (1, 2). These lesions are predominantly located at the corticomedullary junction, characterized by insidious onset, rapid progression, and later manifestations including intracranial hypertension, neurological dysfunction, and epilepsy (3, 4). BM typically indicate advanced disease stage, and their incidence rises with prolonged patient survival. Consequently, early and precise detection is crucial for improving prognosis.

Current BM diagnosis relies heavily on neuroimaging, with magnetic resonance imaging (MRI) serving as the preferred modality due to its lack of ionizing radiation, superior soft-tissue resolution, and multi-sequence capabilities. Compared to computed tomography (CT), MRI demonstrates greater sensitivity for detecting posterior fossa lesions, multiple punctate metastases, and leptomeningeal disease. The typical MRI presentation is a ring-enhancing lesion on contrast-enhanced T1-weighted imaging (CE-T1WI) accompanied by significant peritumoral edema. However, traditional manual identification of multiple (especially small) metastatic foci is time-consuming and carries a high risk of missed diagnosis. Achieving efficient and accurate BM identification therefore remains a significant clinical challenge.

The advancement of artificial intelligence (AI) and radiomics in brain imaging critically depends on voxel-level image segmentation technology (5, 6). This technique partitions image regions based on features like intensity, shape, and texture to integrate targets, forming a fundamental prerequisite for computer-aided image analysis. The U-Net model, introduced by Ronneberger et al. (7), represents a major advancement. It efficiently utilizes limited annotated data, balances localization accuracy with contextual information, and offers advantages such as rapid segmentation, capacity for large image processing, and strong generalization. However, U-Net exhibits limitations, including an output size smaller than the input, dependence on specific tile sizes, constrained applicability of data augmentation techniques, and the requirement for manual loss function parameter tuning. To overcome the constraints of 2D processing, Cicek et al. (8) developed 3D U-Net. This architecture directly learns from sparsely annotated volumetric data to achieve dense 3D segmentation, supporting both semi-automatic and fully automatic workflows. By incorporating batch normalization and weighted loss functions, 3D U-Net significantly enhances performance while retaining the advantages of handling large datasets and robust generalization.

Previous research has developed various computer-aided diagnosis (CAD) systems for BM detection on MRI using diverse algorithms and sequences (9–13). Cho SJ et al. (10) conducted a comparative analysis of 12 recent studies, concluding that deep learning (DL) achieves BM detection rates comparable to classical machine learning approaches, with a lower per-case false positive rate. Despite ongoing CAD development, widespread clinical adoption faces hurdles. Most prior studies on BM detection rates are single-center retrospective analyses (9, 14–18), with the exception of a multicenter retrospective study by Xu J et al. (13). This reliance limits comprehensive evaluation of algorithmic stability and introduces potential selection bias. Furthermore, while previous models predict BM using MRI data (10, 16, 19), their robustness requires more thorough assessment.

This study aims to develop a 3D U-Net deep learning model based on the ResNet-34 backbone network. Through a systematic preprocessing pipeline and a multi-dimensional validation strategy, the model will achieve robust automatic segmentation of BM. Utilizing both internal and external datasets, stratified validation (by gender/age/cancer subtype subgroups) and model robustness testing will be conducted. Concurrently, the lesion detection performance between radiologists and the novel deep learning model for BM will be evaluated. The ultimate goal is to build an AI clinical decision-support tool to enhance both the precision and efficiency of brain tumor imaging diagnosis.

3 Manuscript formatting

4 Materials and methods

4.1 Study design and participants

This retrospective study was approved by the medical ethics committee and the patients' informed consent was waived. A total of 470 patients diagnosed with BM in Jiangsu Cancer Hospital (Nanjing, China) from April 2022 to December 2024 included in our study were divided into 80% training set (379 cases) and 20% internal validation set (91 cases) using random sampling. In addition,172 patients diagnosed with BM at the Affiliated Drum Tower Hospital of Nanjing University Medical School (Nanjing, China) from February 2022 to September 2022 were used for external validation.

Participants who met following criteria were included in this study: (1) Patients were confirmed by clinical examinations to have brain metastases and had completed enhanced MRI scans of the brain. (2) Patients were aged 18 years or older and had complete clinical data. (3) The obtained MR images of patients were free of artifacts and distortion and had relatively high resolution. Meanwhile, to ensure the quality of the study, patients who meet any of following criteria would be excluded: (1) Patients with critical conditions and unstable vital signs. (2) Patients who were unable to tolerate MRI examinations and had only completed plain MRI scans without being able to undergo enhanced scans. (3) Patients with other serious cardiovascular and cerebrovascular diseases. (4) Patients with contraindications for MRI examinations, such as those with implanted cardiac pacemakers. (5) Patients whose imaging data have problems such as severe artifacts, noise, or motion blur (Figure 1).

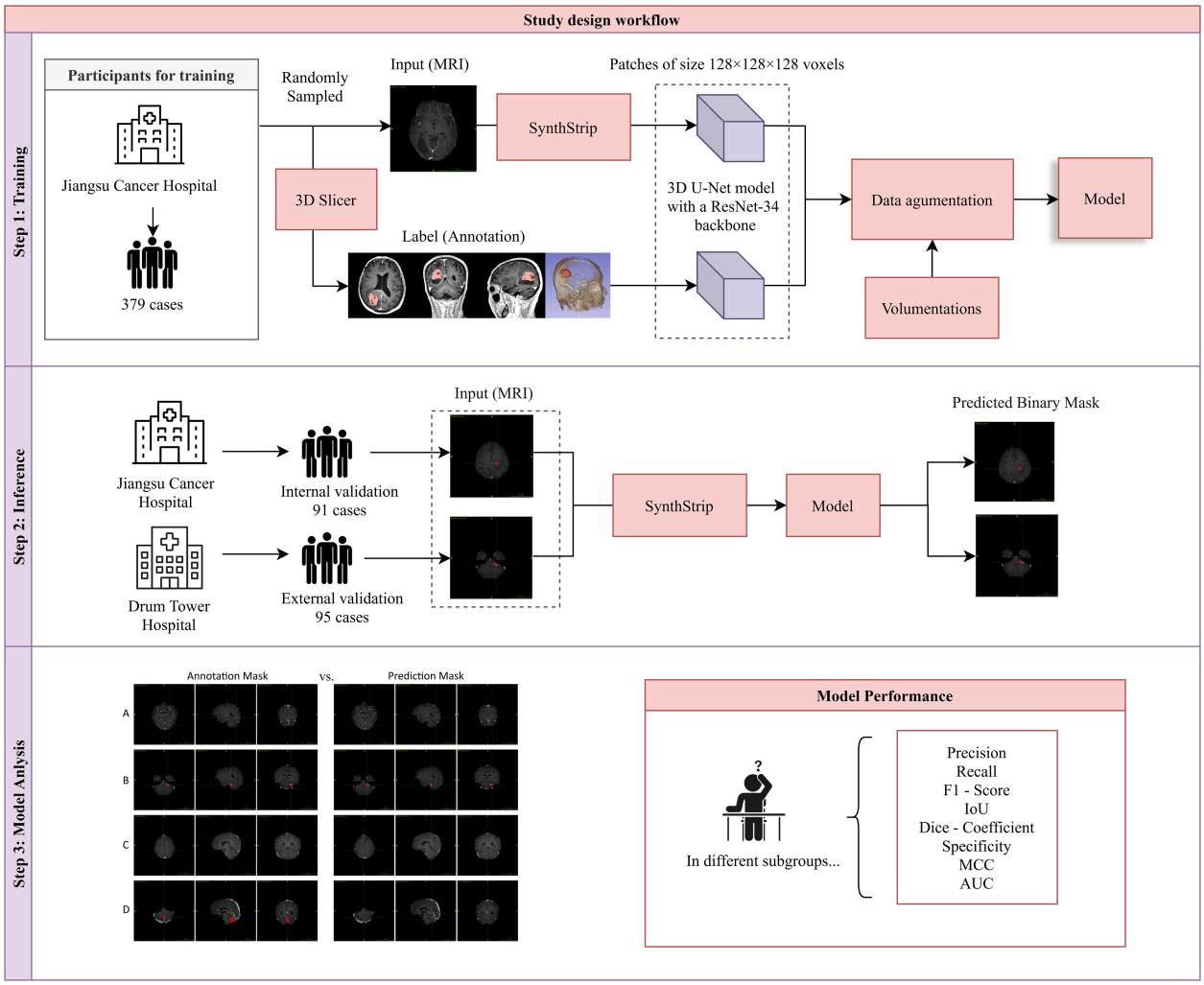

FIGURE 1
Study design workflow. Our study comprised three sequential phases: (1) Model Training: A 3D U-Net model with a ResNet-34 backbone was trained on the training dataset. (2) Inference: The model's performance was evaluated using an internally curated testing set from the same hospital and an external validation set from another hospital. (3) Model Evaluation: Eight metrics were calculated to assess model performance, and a comparative analysis between model-based and manual delineations was conducted.

4.2 MRI contrast-enhanced scan

The 3.0T superconducting MRI machine produced by Philips is selected. Instruct the patient to take the supine position and keep the head stable. A 32-channel head quadrature coil is used to collect data. The contrast injection method is to inject gadopentetate dimeglumine (Gd - DTPA) through the cubital vein at an injection flow rate of 1 mL/s and a dose of 0.25 mL/kg. Scanning and collecting data were performed after a delay of 3 minutes. The sequence settings for the contrast-enhanced 3D - T1WI scan were as follows: the sequence was 3D - FFE, the matrix was 256×256, the slice spacing/slice thickness was 1 mm/0 mm, the TR/TE (ms) was 6.6/3.0, the FOV was 240×240, and the voxel size was 1 mm * 1 mm * 1 mm.

4.3 Image selection and delineation

Three magnetic resonance physicians with more than 8 years of work experience reviewed the films independently, evaluated all the metrics of participants, and selected the magnetic resonance images that met the MRI imaging characteristics of brain metastases. They then manually outlined the enhanced metastatic lesions using software. The enhancement patterns could be divided into standardized uniform nodular enhancement, ring-shaped enhancement, and irregular enhancement. Standardized uniform nodular enhancement described that during the contrast-enhanced scan, both the interior and the edge of the nodular lesion shown uniform and significant enhancement. Ring-shaped enhancement meant that the lesion appeared nodular. During the contrast-enhanced scan, the central part of the lesion was not enhanced, and the enhancement of the edge was incomplete, forming a ring shape. Irregular enhancement was manifested as follows: although the lesion had nodular features, the enhancement effect was not uniform, or the shape of the lesion itself was irregular, and the enhancement patterns were diverse, which might be partially uniform and partially non-uniform.

5 Method

5.1 Implementation

To ensure robust and high-quality processing for model training and evaluation, we implemented a comprehensive preprocessing pipeline and developed a 3D U-Net model with a ResNet-34 backbone tailored for volumetric medical image segmentation. The preprocessing steps focused on standardizing the spatial resolution of MRI scans, normalizing intensity distributions, and implementing quality control measures. The 3D U-Net model utilized a pretrained encoder for efficient feature extraction, a patch-based training strategy to handle large MRI volumes, and data augmentation techniques to enhance model robustness and generalizability (20) (Figure 2).

5.2 Training data

The training data consisted of MRI scans paired with segmentation masks to enable supervised learning for the segmentation task. Preprocessing began with resampling the MRI scans to a uniform voxel size of 0.833 mm^3. This standardization minimized variability caused by differences in scanner resolution and acquisition parameters. An automated verification step was employed to ensure alignment between the MRI scans and their corresponding segmentation masks, flagging mismatched pairs for manual review. Brain extraction was performed using SynthStrip (17), which removed non-brain tissues and reduced computational overhead by focusing on relevant brain regions. Intensity normalization followed, employing Z-score normalization to standardize voxel intensities, mitigating inter-subject variability due to scanner settings or patient-specific factors. Quality control was conducted by visualizing intensity histograms using matplotlib, ensuring consistency and reliability in the preprocessing pipeline.

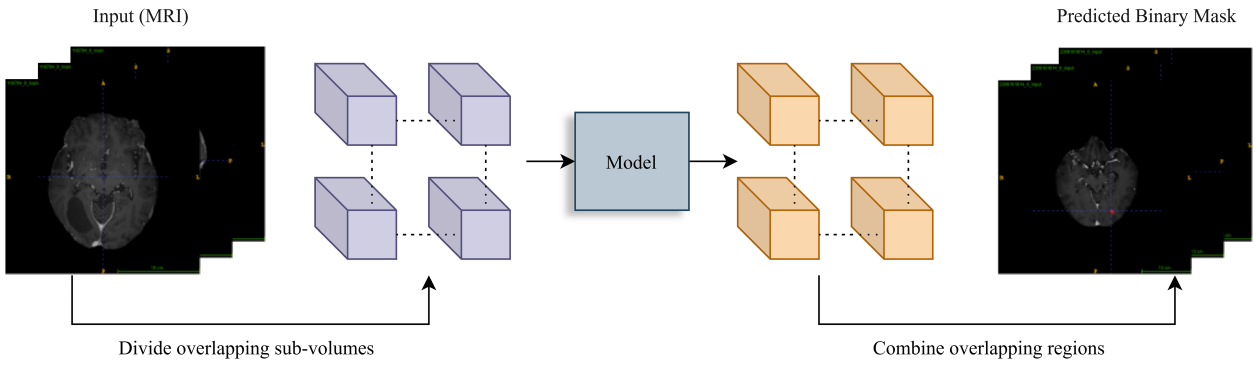

FIGURE 2
Inference workflow. A sliding window (stride=64) divides the input MRI volume into overlapping sub-volumes. Each is processed by the trained model for prediction. Overlapping region outputs are averaged to reconstruct the final binary tumor segmentation mask, ensuring smooth transitions.

5.3 Inference data for generating segmentation masks

The inference data comprised both internal and external datasets. The internal test set included 91 cases drawn from the same distribution as the training data, providing a baseline for performance assessment under known conditions. The external validation set consisted of 95 cases from an independent dataset, offering a robust evaluation of the model's generalizability. Ground-truth segmentation masks were provided for both datasets, enabling a direct comparison of predicted and actual segmentations. This multi-faceted evaluation ensured that the model's performance was rigorously assessed across diverse scenarios.

6 Experiment

6.1 Training

The 3D U-Net model with a ResNet-34 backbone was optimized for volumetric segmentation. The encoder, initialized with pretrained ImageNet weights, accelerated convergence and enhanced feature extraction efficiency. A patch-based training strategy was employed, dividing MRI scans into patches of size 128×128×128 voxels to accommodate the memory constraints posed by the large data volumes. Data augmentation was applied using volumentations, incorporating techniques such as left-right flipping, elastic deformations, and gamma adjustments. These augmentations simulated anatomical and imaging variability, boosting the model's robustness to unseen data. The hybrid loss function, which combined Dice loss and binary cross-entropy with equal weighting, was chosen to balance pixel-wise accuracy with overlap-based evaluation metrics. Optimization was carried out using the Adam optimizer with a learning rate of 0.0001 and parameters $\beta1 = 0.9$, $\beta2 = 0.999$, $\in=1\times10-7\beta1=0.9$, $\beta2=0.999$, $\in=1\times10-7$. To prevent overfitting, a dropout rate of 0.5 was applied in the decoder layers. A batch size of 4 was used, constrained by the memory limitations of the NVIDIA RTX 4090 GPU with 64 GB of system RAM. The training process was implemented in TensorFlow and Keras.

6.2 Inference: generating segmentation masks

The trained 3D U-Net model was applied to both internal and external datasets to generate segmentation masks. MRI scans were preprocessed to align with the model's input requirements, including brain extraction and normalization. A patch-based inference method was adopted, employing overlapping patches with a 50% overlap (stride = 64 voxels). Predictions from overlapping patches were averaged to minimize boundary artifacts and produce smoother segmentation results. Binary segmentation masks were generated by thresholding the model's outputs at 0.5. Performance metrics, including Area Under the Curve (AUC), Dice coefficient, F1-score, Intersection over Union (IoU), Matthew's correlation coefficient (MCC), precision, recall, and specificity were calculated using scikit-learn. These metrics were aggregated and analyzed using pandas for statistical analysis.

6.3 Statistical analysis

For the descriptive analysis, continuous variables with normal distribution were presented as mean and standard deviation (SD), and those with a non-normal distribution were presented as median and interquartile range (IQR). Categorical variables were summarized as counts (n) with corresponding percentages (%). Continuous variables were assessed for statistical differences using two-sample t test or Mann–Whitney U tests. Categorical variables were evaluated for differences using the $\chi2$ test. Subgroup analyses were conducted to identify the population in which this model is most suitable for delineating lesions between each subgroup. The internal testing set and the external validation set were both stratified by age (< 60 years and >= 60 years), gender, and primary cancer type (lung cancer and other cancer). For metrics (precision, recall, F1 score, IoU, dice coefficient, specificity, and MCC), differences between each subgroup were assessed by employing Mann–Whitney U tests, with corresponding 95% confidence intervals (CIs) estimated. The comparison of AUC between subgroups was performed by using DeLong test. Additionally, the differences of time cost between model-based delineation and manual annotation were also compared by using Mann–Whitney U tests in both internal testing set and external validation set. All statistical analyses were performed using R version 4.4.1. A P-value of less than 0.05 was considered statistically significant.

7 Result

7.1 Patient characteristics

This retrospective study included 91 patients in the test set and 172 patients in the validation set. The median baseline ages in the test and validation sets were 59.4 ± 10.2 and 57.4 ± 12.6 years, respectively. Demographic and clinical characteristics, including age and gender, were comparable between the two cohorts, with no statistically significant differences observed. Selection bias was present regarding cancer type, showing a statistically significant difference. The clinical characteristics of the patients are summarized in Table 1.

7.2 Model performance evaluation

Table 2 describes the case characteristics in the internal test set and external validation set. In our study, model performance on both the internal test set and external validation set was evaluated by calculating Precision, Recall, F1 Score, Intersection over Union

TABLE 1 Characteristics of patients in testing set and validation set.

Characteristics	Testin set (*N* = 91)	Validation set (*N* = 95)	*p*-value
Age	59.4 ± 10.2	57.4 ± 12.6	0.173
<60	49	82	0.411
≥60	42	90	
Gender			0.136
Male	5	84	
Female	37	88	
Cancer type			<0.001
Lung cancer	74	93	
Other type	17	79	

(IoU), Dice Coefficient, Specificity, Matthews Correlation Coefficient (MCC), and the Area Under the Receiver Operating Characteristic Curve (AUC). Results were summarized using median and interquartile range (IQR) (Table 2). The results showed that the model achieved an AUC of 0.89 (IQR 0.79–0.93) on the internal test set and 0.82 (IQR 0.67–0.90) on the validation set, indicating good overall diagnostic capability. Furthermore, Precision was >0.93 in both the internal test set and external validation set, suggesting strong discriminative ability, with the external validation set showing higher precision than the test set. Recall decreased from 0.78 (IQR 0.57–0.86) in the test set to 0.64 (IQR 0.34–0.81) in the validation set, accompanied by a synchronous decline in F1 Score (0.82 vs. 0.75), indicating moderate model robustness. Specificity reached 1.000 (IQR 1.000–1.000) in both the test and validation sets, signifying theoretically optimal ability to exclude negative samples. MCC decreased to 0.77 (IQR 0.55–0.86) in the validation set, indicating good classification reliability; the Dice Coefficient also decreased synchronously to 0.75 (IQR 0.49–0.85).

As shown in Figure 3a comparison between clinicians and the model in identifying brain metastatic lesions was conducted, both for the test set and the validation set. The clinical validation data indicated that compared with the manual interpretation by clinicians, the number of lesions detected by this model was significantly higher, effectively reducing the rate of missed detections, highlighting its potential value in improving the accuracy of clinical diagnosis.

TABLE 2 Summary descriptive table by groups of datasets.

Metrics	Testing set (*N* = 91)	Validation set (*N* = 172)
Precision	0.926(0.831,0.966)	0.935(0.858,0.967)
Recall	0.782(0.567,0.861)	0.638(0.342,0.806)
F1 Score	0.821(0.674,0.882)	0.750(0.495,0.852)
IoU	0.697(0.508,0.789)	0.600(0.329,0.743)
Dice Cofficient	0.821(0.674,0.882)	0.750(0.495,0.852)
Specificity	1.000(1.000,1.000)	1.000(1.000,1.000)
MCC	0.832(0.698,0.883)	0.768(0.549,0.857)
AUC	0.891(0.793,0.930)	0.819(0.671,0.903)

7.3 Model performance comparative analysis

To comprehensively evaluate model robustness, this study conducted stratified validation across three dimensions (gender, age, and primary cancer type). Additionally, stability assessments were performed on both datasets (internal test set and external independent validation set).

Figure 4 systematically presents the multi-dimensional evaluation results of the model in the internal test set and the external validation cohort.

In the internal test set (n = 91): Our model demonstrated superior predictive performance in males, patients < 60 years old, and patients with other cancer compared with the other groups. The AUC was 0.89 for males, 0.91 for patients younger than 60 years old, and 0.91 for patients with other cancer type. DeLong tests indicated that these between-group differences were statistically significant (all $p < 0.001$). Subgroup analysis by cancer type (lung cancer n = 74, other cancers n = 17) revealed: There was no significant difference of other metrics between groups ($p > 0.05$), and its high specificity (1.0000) and stable AUC (> 0.88) supported generalization ability. The median of other cancer group comprehensive indicators (F1/IoU/MCC) was higher, suggesting that the model may have better discrimination ability for non-lung cancer malignancies. Subgroup analysis by gender (male n = 54, female n = 37) showed that apart from the AUC, the differences of other performance metrics between subgroups were not statistically significant ($p > 0.10$), and the model performance was highly consistent; Specificity reached 1.0000, indicating that the model's ability to exclude false positives was extremely strong. Age subgroup analysis (< 60 years old n = 49, ≥ 60 years old n = 42) indicated: The age factor did not significantly change the model performance metrics ($p > 0.20$); The lower limit of the recall rate distribution in the ≥ 60 years old group (Q1 = 0.5540) suggested that the sensitivity needed to be optimized to reduce the risk of missed diagnosis in the elderly.

In the external validation set (n = 172), the prediction model achieved better comprehensive performance in the male subgroup (AUC = 0.86), the subgroup aged ≥60 years (AUC = 0.90), and the subgroup with other cancers (AUC = 0.89). DeLong tests indicated that the differences of AUC between subgroups were statistically significant ($p < 0.001$). The subgroup analysis by cancer type (lung cancer n = 93, other cancers n = 79) showed that the precision of the other cancer group was significantly higher ($p < 0.001$), indicating that the model has a discriminative advantage in identifying non-lung cancer malignancies. Both subgroups maintained a perfect specificity (1.0000), verifying the model's generalization ability in controlling false positives; the differences in the remaining indicators between the lung cancer group and the non-lung

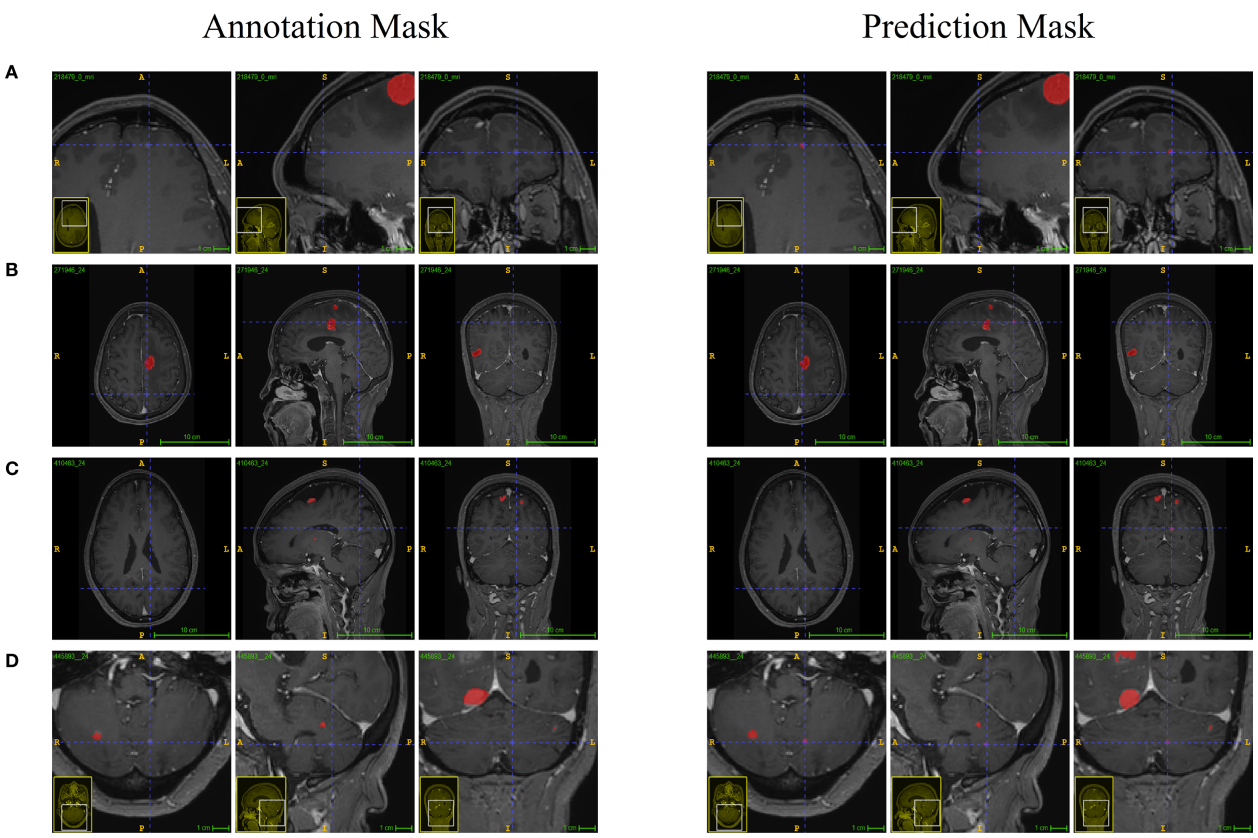

FIGURE 3
Manual identification and model identification of the lesion legend.

cancer group were not statistically significant (all $p > 0.26$), suggesting that the model performance is cancer type-dependent. The comprehensive performance of the lung cancer group (AUC = 0.82) was significantly lower than that of the other cancer group, indicating that the model needs to be further optimized for its sensitivity to lung cancer. The statistical analysis grouped by gender (male n = 84; female n = 88) showed that except for AUC, the differences in evaluation indicators between the male group and the female group were not statistically significant (all $p > 0.52$), indicating that the model performance is stable across the gender dimension. The high specificity (1.0000) and stable AUC value (> 0.84) remained consistent in the cross-center validation, reflecting the model's good robustness. In the age subgroup analysis (< 60 years old n = 82, ≥ 60 years old n = 90), apart from AUC, the differences in other indicators between the two age groups were not statistically significant ($p > 0.05$), indicating that the model performance was highly consistent across different age groups; the perfect specificity (1.0000) and stable AUC value (> 0.82) further confirmed the reliability of the model in excluding false positives.

To further demonstrate the clinical practicability of this model, we conducted a comparative analysis of the time spent on model-based depiction and manual lesion annotation. The results showed that, in both datasets, the time required for model-based depiction was significantly less than that of manual annotation (both $P < 0.001$) (Table 3), and it maintained a comparable consistency with manual depiction (Figure 3). In the internal test set, the median time of the model was 69s (IQR 68 - 69), while the median time of manual annotation was 113 seconds (IQR 90 - 151). In the external validation set, the median time of the model was 66s (IQR 65 - 68), while the median time of manual annotation was 96s (IQR 78 - 113).

8 Discussion

In this multicenter study, we developed a 3D U-Net deep learning model based on the ResNet-34 backbone network for the detection of brain metastases (BM) in 3D CET1WI MRI images of a large number of patients. Additionally, we conducted a multi-reader evaluation aimed at quantifying the impact of the BM-assisted system on reading time and lesion detection efficacy. Through multi-dimensional stratified validation, this study systematically evaluated the robustness of the model on the internal test set and the external independent validation set. The results showed that the model demonstrated perfect specificity (Specificity = 1.0000) for both male and female patients, strongly confirming its ability to exclude false positive results regardless of gender factors. Similarly, in the age stratification analysis (<60 years vs. ≥60 years), the core efficacy indicators (such as AUC, specificity) were highly consistent between the two age groups, supporting its universality in patients of different ages. In the cancer type dimension, the model performed comparably

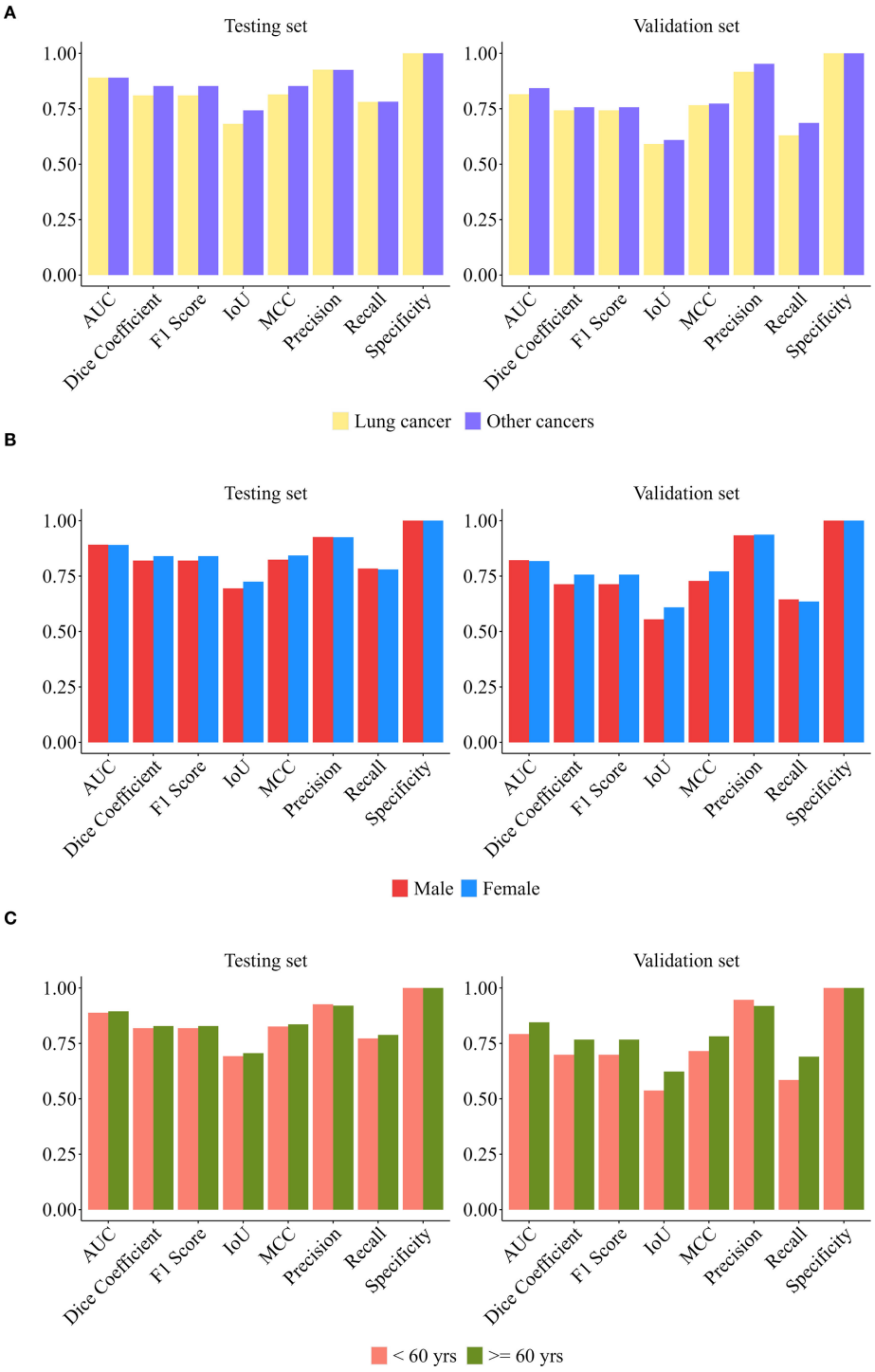

FIGURE 4
Training vs. validation set comparisons by subgroup. **(A)** Cancer type **(B)** Gender **(C)** Age.

in the internal test set for lung cancer and other malignant tumors overall; while in the external validation set, it showed a significantly higher precision rate for non-lung cancer malignant tumors (the "other cancer" group). Although both groups maintained perfect specificity, the comprehensive performance indicators of the lung cancer group in the external validation were relatively lower, suggesting that the model performance may have some cancer type dependence. The model yielded better predictive performance when applied to the subgroups of males and non-lung cancer patients. Future optimization should focus on improving the model's sensitivity for detecting lung cancer lesions. Overall, this model significantly improved the efficiency in detecting BM (with much less time than manual annotation), and its excellent robustness demonstrated a good potential for clinical translation applications.

TABLE 3 Comparison of time consuming in model-based and manual annotation in internal testing set and external validation set.

Dataset	Time, s		Z	P
	Model	Annotation		
Testing	69 (68,69)	113 (90,151)	-11.672	<0.001
Validation	66 (65, 68)	96 (78,113)	-14.143	<0.001

BMs are the most common intracranial malignant tumors, with an incidence of approximately 20%-40% (1, 4, 21). BM often leads to severe neurological lesions and shortens survival time, and its early diagnosis is closely related to clinical decisions and patient prognosis (22). MRI, with its high soft tissue resolution and rich scanning sequences, has become the main examination method for BM. Although MRI can provide various imaging features, manual visual diagnosis often fails to cover all effective features and ignores the complexity of tissue cells, resulting in difficult improvement of detection accuracy to the application level. Therefore, developing an automated brain tumor segmentation technology to achieve high-precision and repeatable measurement of tumor substructures, replacing the current manual basic assessment, has become an urgent need to supplement BM diagnosis. Fajam et al. (9) in a prospective single-center study included 29 patients and developed a set of uneven 3D spherical template to detect brain metastases in Gd-enhanced T1WI imaging, achieving a sensitivity of 93.5% and an intracranial false positive rate of 0.024. However, the system in their study was not evaluated clinically. Lu SL et al. (12) demonstrated through a random multi-reader multi-case study of 10 patients (a total of 23 tumors) that the automatic detection and segmentation technology based on deep neural networks (ABS system) can improve the accuracy and efficiency of tumor contour delineation and reduce the differences among doctors. However, this study focused more on tumor segmentation rather than detection, and did not analyze reading time. Xue J et al. (13) retrospectively included 1625 patients from three centers and constructed a model for detecting and segmenting brain metastases, although the sensitivity, specificity, and Dice coefficient of the model were evaluated, the reading time of the model compared with manual recognition was not assessed. Unlike previous studies, this research employs a 3D U-Net deep learning model based on the ResNet-34 backbone to detect brain metastases on 3D enhanced T1-weighted MRI images of a large number of patients. Through multi-dimensional hierarchical validation on the internal test set and external validation set, the BM segmentation precision of this 3D U-Net model reaches 92.6% and 93.5% respectively. Meanwhile, the specificity of our model in the internal test set and external validation set is also very outstanding. Compared with previous studies, the model in this research has higher precision and specificity (1, 23). Amemiya S et al. developed a combined algorithm based on feature fusion and single detector (24), which has a high overall sensitivity and specificity for brain metastases without reducing the positive predictive value, thereby improving the detection rate of small lesions. Huang et al. (25) proposed a deep learning model based on volume-level sensitivity and specificity, which also shows high sensitivity and accuracy for BM detection.

One of the most commonly used network architectures is the so-called U-Net (26, 27). Recently, Pfluger I et al. developed a system based on artificial neural networks (28) using the nnU-Net method to segment meningiomas from 308 patients (29). Bousabarah K et al. implemented traditional U-Net and an improved U-Net with multiple outputs to achieve automatic separation of meningiomas (30). Additionally, Gong J et al. proposed an integrated learning model based on deep learning and image-based radiomics to improve the prediction ability of the risk of brain metastasis in patients with advanced non-small cell lung cancer within three years (31). By applying the deep residual U-Net model, each pulmonary tumor can be automatically and accurately segmented, combined with CT image-based radiomics and clinical features. This improves the performance of predicting the risk of meningiomas. We also found that the quality of automatic segmentation is very high ($0.8 \geq DSC \geq 0.6$) in both the internal test set and external validation set, indicating that the model can provide more reliable image segmentation for brain metastases. Zhu Genste et al. trained a deep learning model for detecting and three-dimensional segmenting brain metastases in non-small cell lung cancer (32). Compared with manual segmentation, the DSC consistency coefficient reached 0.72.

Most previous studies used classical machine learning or deep learning and multiple sequences to detect or segment brain metastases based on the number and size of the lesions (33, 34). Although these studies mostly evaluated overall sensitivity, accuracy, and per capita false positive rate, they covered various different primary tumor subtypes. Although multi-modal modalities have significant advantages, the additional scanning time and sequence availability costs may hinder their widespread clinical application. Our method uses only the CE TIWI sequence and can detect brain metastases with a precision of over 80%. Given the trade-off between precision and specificity, our model may be more suitable for wide clinical application. Unlike previous studies, this research validates the model through multi-dimensional hierarchical validation, firmly confirming that the model is not affected by factors such as gender and age, supporting its universality in patients of different ages and genders. Although this model shows certain type-dependent performance in the cancer type dimension for the identification of lung cancer and other cancers, its efficiency has significantly improved and the overall performance is quite good, demonstrating good potential for clinical translation applications.

Our research indicates that the time required for lesion identification based on the model is significantly shorter than that of manual identification, effectively improving the efficiency of clinical diagnosis. Compared with other technologies (33, 34), this study has several unique features. Firstly, our research includes training sets and validation sets to verify our system, which is more persuasive than previous studies. Secondly, to evaluate the stability of the model, this study conducts stratified validation from multiple dimensions (gender, age, and cancer type), and the results show that the internal and external validation sets of the model exhibit excellent robustness, showing good universality for patients of different genders and different age groups. The model has a certain type-dependent performance for cancer. Moreover, the efficiency of this model in identifying BM is significantly higher than manual

annotation. Combined with its excellent robustness, it has good potential for clinical translation applications.

This study also has potential limitations. Firstly, the limiting factor for the model to achieve higher maturity is the small number of samples of BM patients included in the training and validation sets. Secondly, our model has some false negative results in identifying BM; BM and blood vessels may appear as nodules or high-signal spots on CE T1WI, and the latter may be mistakenly regarded as BM. Finally, it is necessary to validate our model on a larger dataset and further verify its stability in multiple centers. Subsequently, a stratified analysis of lesion size will be conducted to prevent missed diagnoses of lesions.

In summary, our study shows that the 3D U-Net model demonstrates perfect specificity and robustness across transgender and age groups, with slight differences in cancer types (the sensitivity of lung cancer needs to be optimized). Overall, it is highly efficient and accurate, and has significant clinical translational value. Based on our multi-center evaluation, this system helps radiologists with different levels of experience achieve higher detection specificity and precision.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Medical Ethics Committee of Jiangsu Cancer Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because This study is a retrospective investigation for which a waiver of the patient informed consent form was obtained.

Author contributions

DS: Investigation, Writing – original draft. MY: Writing – review & editing, Writing – original draft. MD: Writing – review & editing. NX: Writing – original draft, Methodology, Software. YZ: Funding acquisition, Writing – original draft. XL: Writing – original draft, Conceptualization. CX: Validation, Writing – original draft. FX: Writing – original draft, Resources. LX: Supervision, Writing – review & editing, Methodology. QZ: Supervision, Writing – review & editing, Project administration. NY: Writing – review & editing, Data curation, Writing – original draft.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. The research was supported by General Program of Jiangsu Provincial Natural Science Foundation of China (grant number: BK20231369).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1657604/full#supplementary-material

References

1. Lamba N, Wen PY, Aizer AA. Epidemiology of brain metastases and leptomeningeal disease. *Neuro Oncol.* (2021) 23:1447–56. doi: 10.1093/neuonc/noab101

2. Boire A, Brastianos PK, Garzia L, Valiente M. Brain metastasis. *Nat Rev Cancer.* (2020) 20:4–11. doi: 10.1038/s41568-019-0220-y

3. Vogelbaum MA, Brown PD, Messersmith H, Brastianos PK, Burri S, Cahill D, et al. Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline [published correction appears in J Clin Oncol. *2022 Apr 20;40(12):1392. doi: 10.1200/JCO.22.00593.*]. *J Clin Oncol.* (2022) 40:492–516. doi: 10.1200/JCO.21.02314

4. Suh JH, Kotecha R, Chao ST, Ahluwalia MS, Sahgal A, Chang EL. Current approaches to the management of brain metastases. *Nat Rev Clin Oncol*. (2020) 17:279–99. doi: 10.1038/s41571-019-0320-3

5. Raj GM, Dananjayan S, Gudivada KK. Applications of artificial intelligence and machine learning in clinical medicine: What lies ahead? *Med Adv*. (2024) 2:202–4. doi: 10.1002/med4.62

6. Wen Q, Qiu L, Qiu C, Che K, Zeng R, Wang X, et al. Artificial intelligence in predicting efficacy and toxicity of Immunotherapy: Applications, challenges, and future directions. *Cancer Lett*. (2025) 630:217881. doi: 10.1016/j.canlet.2025.217881

7. Ronneberger O, Fischer P, Brox T. (2015). U-net: convolutional networks for biomedical image segmentation, in: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Lecture Notes in Computer Science: Springer, Cham. 9351:234–241. doi: 10.1007/978-3-319-24574-4_28

8. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016). 3D U-net: learning dense volumetric segmentation from sparse annotation, in: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W. (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Lecture Notes in Computer Science: Springer, Cham. 9901:424–32. doi: 10.1007/978-3-319-46723-8_49

9. Farjam R, Parmar HA, Noll DC, Tsien CI, Cao Y. An approach for computer-aided detection of brain metastases in post-Gd T1-W MRI. *Magn Reson Imaging*. (2012) 30:824–36. doi: 10.1016/j.mri.2012.02.024

10. Cho SJ, Sunwoo L, Baik SH, Bae YJ, Choi BS, Kim JH. Brain metastasis detection using machine learning: a systematic review and meta-analysis. *Neuro Oncol*. (2021) 23:214–25. doi: 10.1093/neuonc/noaa232

11. Park YW, Jun Y, Lee Y, Han K, An C, Ahn SS, et al. Robust performance of deep learning for automatic detection and segmentation of brain metastases using three-dimensional black-blood and three-dimensional gradient echo imaging. *Eur Radiol*. (2021) 31:6686–95. doi: 10.1007/s00330-021-07783-3

12. Lu SL, Xiao FR, Cheng JC, Yang WC, Cheng YH, Chang YC, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. *Neuro Oncol*. (2021) 23:1560–8. doi: 10.1093/neuonc/noab071

13. Xue J, Wang B, Ming Y, Liu X, Jiang Z, Wang C, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. *Neuro Oncol*. (2020) 22:505–14. doi: 10.1093/neuonc/noz234

14. Zhou Z, Sanders JW, Johnson JM, Gule-Monroe MK, Chen MM, Briere TM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. *Radiology*. (2020) 295:407–15. doi: 10.1148/radiol.2020191479

15. Ambrosini RD, Wang P, O'Dell WG. Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching. *J Magn Reson Imaging*. (2010) 31:85–93. doi: 10.1002/jmri.22009

16. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput Biol Med*. (2018) 95:43–54. doi: 10.1016/j.compbiomed.2018.02.004

17. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Magn Reson Imaging*. (2020) 51:175–82. doi: 10.1002/jmri.26766

18. Cao P, Jia X, Wang X, Fan L, Chen Z, Zhao Y, et al. Deep learning radiomics for the prediction of epidermal growth factor receptor mutation status based on MRI in brain metastasis from lung adenocarcinoma patients. *BMC Cancer*. (2025) 25:443. doi: 10.1186/s12885-025-13823-8

19. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M. SynthStrip: skull-stripping for any brain image. *Neuroimage*. (2022) 260:119474. doi: 10.1016/j.neuroimage.2022.119474

20. Bonada M, Rossi LF, Carone G, Panico F, Cofano F, Fiaschi P, et al. Deep learning for MRI segmentation and molecular subtyping in glioblastoma: critical aspects from an emerging field. *Biomedicines*. (2024) 12:1878. doi: 10.3390/biomedicines12081878

21. Zhang M, Young GS, Chen H, Li J, Qin L, McFaline-Figueroa JR, et al. Deep-learning detection of cancer metastases to the brain on MRI. *J Magn Reson Imaging*. (2020) 52:1227–36. doi: 10.1002/jmri.27129

22. Ghaderi S, Mohammadi S. Utility of ultra-high-field magnetic resonance imaging in the detection and management of brain metastases. *Med Adv*. (2024) 2:199–201. doi: 10.1002/med4.59

23. Dikici E, Ryu JL, Demirer M, Bigelow M, White RD, Slone W, et al. Automated brain metastases detection framework for T1-weighted contrast-enhanced 3D MRI. *IEEE J BioMed Health Inform*. (2020) 24:2883–93. doi: 10.1109/JBHI.2020.2982103

24. Amemiya S, Takao H, Kato S, Yamashita H, Sakamoto N, Abe O. Feature-fusion improves MRI single-shot deep learning detection of small brain metastases. *J Neuroimaging*. (2022) 32:111–9. doi: 10.1111/jon.12916

25. Huang Y, Bert C, Sommer P, Frey B, Gaipl U, Distel LV, et al. Deep learning for brain metastasis detection and segmentation in longitudinal MRI data. *Med Phys*. (2022) 49:5773–86. doi: 10.1002/mp.15863

26. Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput Biol Med*. (2021) 136:104699. doi: 10.1016/j.compbiomed.2021.104699

27. Punn NS, Agarwal S. Modality specific U-Net variants for biomedical image segmentation: a survey. *Artif Intell Rev*. (2022) 55:5845–89. doi: 10.1007/s10462-022-10152-1

28. Pflüger I, Wald T, Isensee F, Schell M, Meredig H, Schlamp K, et al. Automated detection and quantification of brain metastases on clinical MRI data using artificial neural networks. *Neurooncol Adv*. (2022) 4:vdac138. doi: 10.1093/noajnl/vdac138

29. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

30. Bousabarah K, Ruge M, Brand JS, Hoevels M, Rueß D, Borggrefe J, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. *Radiat Oncol*. (2020) 15:87. doi: 10.1186/s13014-020-01514-6

31. Gong J, Wang T, Wang Z, Chu X, Hu T, Li M, et al. Enhancing brain metastasis prediction in non-small cell lung cancer: a deep learning-based segmentation and CT radiomics-based ensemble learning model. *Cancer Imaging*. (2024) 24:1. doi: 10.1186/s40644-023-00623-1

32. Jünger ST, Hoyer UCI, Schaufler D, Laukamp KR, Goertz L, Thiele F, et al. Fully automated MR detection and segmentation of brain metastases in non-small cell lung cancer using deep learning. *J Magn Reson Imaging*. (2021) 54:1608–22. doi: 10.1002/jmri.27741

33. Pérez-Ramírez Ú, Arana E, Moratal D. Brain metastases detection on MR by means of three-dimensional tumor-appearance template matching. *J Magn Reson Imaging*. (2016) 44:642–52. doi: 10.1002/jmri.25207

34. Sunwoo L, Kim YJ, Choi SH, Kim KG, Kang JH, Kang Y, et al. Computer-aided detection of brain metastasis on 3D MR imaging: Observer performance study. *PloS One*. (2017) 12:e0178265. doi: 10.1371/journal.pone.0178265