Radiotherapy and Oncology 210 (2025) 111031

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Improving risk assessment of local failure in brain metastases patients using vision transformers – A multicentric development and validation study

Ayhan Can Erdur [a,b,*], Daniel Scholz [b,v], Q.Mai Nguyen [a], Josef A. Buchner [a], Michael Mayinger [d], Sebastian M. Christ [d], Thomas B. Brunner [e], Andrea Wittig [f], Claus Zimmer [c], Bernhard Meyer [g], Matthias Guckenberger [d], Nicolaus Andratschke [d], Rami A. El Shafie [h,i,j], Jurgen Debus [h,i], Susanne Rogers [k], Oliver Riesterer [k], Katrin Schulze [l], Horst J. Feldmann [l], Oliver Blanck [m,q], Constantinos Zamboglou [n,o,p], Angelika Bilger-Z. [n,o], Anca L. Grosu [n,o], Robert Wolff [q,r], Kerstin A. Eitz [a,s,t], Stephanie E. Combs [a,s,t], Denise Bernhardt [a,s], Benedikt Wiestler [v], Daniel Rueckert [b,u,1], Jan C. Peeken [a,s,t,1,*]

[a] Department of Radiation Oncology, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany
[b] Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany
[c] Department of Neuroradiology, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany
[d] Department of Radiation Oncology, University of Zurich, Zurich, Switzerland
[e] Department of Radiation Oncology, University Hospital Magdeburg, Magdeburg, Germany
[f] Department of Radiation Oncology, University Hospital Wurzburg, Julius-Maximilians-University, Wurzburg, Germany
[g] Department of Neurosurgery, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany
[h] Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[i] Heidelberg Institute for Radiation Oncology (HIRO), National Center for Radiation Oncology (NCRO), Heidelberg, Germany
[j] Department of Radiation Oncology, University Medical Center Gottingen, Göttingen, Germany
[k] Radiation Oncology Center KSA-KSB, Kantonsspital Aarau, Aarau, Switzerland
[l] Department of Radiation Oncology, General Hospital Fulda, Fulda, Germany
[m] Department of Radiation Oncology, University Medical Center Schleswig Holstein, Kiel, Germany
[n] Department of Radiation Oncology, University of Freiburg - Medical Center, Freiburg, Germany
[o] German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg, Germany
[p] Department of Radiation Oncology, German Oncology Center, European University of Cyprus, Limassol, Cyprus
[q] Saphir Radiosurgery Center Frankfurt and Northern Germany, Kiel, Germany
[r] Department of Neurosurgery, University Hospital Frankfurt, Frankfurt, Germany
[s] Deutches Konsortium fur Translationale Krebsforschung (DKTK), Partner Site Munich, Munich, Germany
[t] Institute of Radiation Medicine (IRM), Helmholtz Center Munich, Munich, Germany
[u] Department of Computing, Imperial College London, London, United Kingdom
[v] AI for Image-Guided Diagnosis and Therapy, TUM School of Medicine and Health, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
Artificial Intelligence
Vision Transformers
Brain metastases
Stereotactic radiotherapy

ABSTRACT

Background and purpose: This study investigates the use of Vision Transformers (ViTs) to predict Freedom from Local Failure (FFLF) in patients with brain metastases using pre-operative MRI scans. The goal is to develop a model that enhances risk stratification and informs personalized treatment strategies.
Materials and methods: Within the AURORA retrospective trial, patients (n = 352) who received surgical resection followed by post-operative stereotactic radiotherapy (SRT) were collected from seven hospitals. We trained our ViT for the direct image-to-risk task on T1-CE and FLAIR sequences and combined clinical features along the way. We employed segmentation-guided image modifications, model adaptations, and specialized patient sampling strategies during training. The model was evaluated with five-fold cross-validation and ensemble learning across all validation runs. An external, international test cohort (n = 99) within the dataset was used to assess the generalization capabilities of the model, and saliency maps were generated for explainability analysis.

* Corresponding authors.
E-mail addresses: can.erdur@tum.de (A.C. Erdur), jan.peeken@tum.de (J.C. Peeken).
[1] Contributed equally as senior authors to this work.

https://doi.org/10.1016/j.radonc.2025.111031
Received 21 March 2025; Received in revised form 17 June 2025; Accepted 2 July 2025
Available online 5 July 2025

Results: We achieved a competent C-Index score of 0.7982 on the test cohort, surpassing all clinical, CNN-based, and hybrid baselines. Kaplan-Meier analysis showed significant FFLF risk stratification. Saliency maps focusing on the BM core confirmed that model explanations aligned with expert observations.
Conclusion: Our ViT-based model offers a potential for personalized treatment strategies and follow-up regimens in patients with brain metastases. It provides an alternative to radiomics as a robust, automated tool for clinical workflows, capable of improving patient outcomes through effective risk assessment and stratification.

Introduction

Brain metastases (BMs) are ten times more common than primary brain tumors and the most prevalent intracranial tumor type. They affect about 20 % of cancer patients and cause significant morbidity and mortality [1,2]. For symptomatic or large BMs, guidelines recommend surgery. Post-operative stereotactic radiotherapy (SRT) to the resection cavity improves local control by achieving up to 70–90 % control rates at twelve months [3,4].

Understanding individual local failure (LF) better can enhance therapy, allowing high-risk patients to benefit from escalated SRT doses, systemic therapies that penetrate the blood–brain barrier, and more frequent imaging follow-ups for early detection of potential failures [5].

In a recent study, Buchner *et al.* achieved a concordance index (C-Index) of 0.77 in external validation for predicting freedom-from-local-failure (FFLF) by integrating radiomic features from contrast-enhancing pre-therapeutic brain metastases (BMs) and surrounding edema with clinical data [5]. Other studies have used radiomic features for FFLF prediction [6–8], often reducing the task to binary predictions at specific time points and without external validation. Radiomics models require complex feature selection based on domain expertise, whereas deep learning models automatically extract features and learn data representations with multiple levels of abstraction, enabling end-to-end training [9].

The advancements in deep learning have impacted medical research as well, with studies exploring brain metastasis local failure prediction on pre-therapeutic MRIs [10–13]. These studies often treat it as a binary prediction problem at a specific time cutoff [11–13] or use a two-stage approach, separating feature extraction from final prediction [10].

While Convolutional Neural Networks (CNNs) have been considered the standard in computer vision, the Vision Transformer (ViT) has recently emerged as a powerful alternative. ViTs, which adapt the state-of-the-art natural language processing (NLP) architecture, transformers to image analysis, have shown superior results to CNNs like ResNet [14,15]. In ViTs, images are divided into patches that serve as input tokens, similar to words in NLP, and fed through the network as *sentences*. The self-attention mechanism among tokens in ViTs allows them to process local and global dependencies across an entire image, improving their ability to understand spatial relationships [14].

Deep learning models are often criticized for their lack of interpretability [16]. However, the attention mechanism in ViTs can be visualized to support interpretability by highlighting the parts to which the network gives importance [17,18].

The objective of this project was to explore the capabilities of ViTs to develop a deep learning model predicting FFLF after resection and SRT of BMs as a right-censored time-to-event endpoint. We trained ViTs on pre-treatment MRI scans and clinical features, using Cox partial log-likelihood to predict hazard ratios for FFLF, adapting survival modeling to deep learning for a direct from-image-to-risk estimation. The models were externally validated with a multicenter international test cohort.

Materials and methods

AURORA study

As part of the retrospective AURORA trial (A Multicenter Analysis of Stereotactic Radiotherapy to the Resection Cavity of Brain Metastases), MR imaging and clinical data were collected from 352 eligible patients across seven hospitals. Inclusion criteria required a known primary tumor, treated with BM resection and subsequent SRT (>5 Gy per fraction) within 100 days after surgery. Patients with prior cranial RT or premature RT discontinuation were excluded. Synchronous non-resected BMs were treated concurrently. See Table 1 for patient characteristics.

LF was determined through radiological assessment at 3-month intervals by a board-certified radiologist or via histology after surgical removal of recurring BMs. FFLF was measured from the end of SRT to LF, with patients censored at their last imaging follow-up if LF did not occur.

Further details on hospitals, treatments, doses, event distribution, and ethical approval are available in supplementary material A.1.

Dataset

The imaging data consisted of four pre-operative scans: T1-weighted, T1-CE (contrast-enhanced), T2-weighted, and T2-FLAIR (fluid-attenuated inversion recovery) sequences. Due to the absence of T2 scans in 34 % of patients, we excluded them. As they contain the highest intensity contrast between healthy and malignant tissue and are rich in information, we used only T1-CE and FLAIR sequences in our study.

Expert-delineated BM core and edema segmentations were available alongside segmentations by the U-Net algorithm developed and tested in the same patient cohort [19].

Tabular clinical data included *age*, *Karnofsky index* (KPS), *primary cancer type*, *BM location*, and binary indicators for *chemotherapy* and *immunotherapy* during primary tumor treatment. To ensure all inputs were pre-therapy, we excluded metastasis-specific treatment details (e.g., total Gy).

For training, 253 patients from two centers (TUM, USZ) were separated from an external, multi-center international test group of 99 patients from five centers (FD, FFM, FR, HD, and KSA). See supplementary material A.1 for details on dataset partitioning.

The same group of patients was used in [5] for FFLF prediction using radiomic features.

Pre-processing

MRI sequences were pre-processed using the BraTS Toolkit [20], yielding co-registered, skull-stripped images with 1 mm isotropic resolution in the SRI-24 atlas (240 × 240 × 155 voxels). Intensities were min–max normalized to [0, 1] based on the 0.5th–99.5th percentiles. For a more focused input, we cropped scans to a volume of interest (VOI) around the edema tissue of the BMs. For patients with multiple BMs (120 in total), we used the largest GTV associated with recorded LF information. Cropping was guided by U-Net segmentations [19], underlining a fully automatic pipeline.

Cropped volumes varied in size, so we standardized them by expanding the cropping boundaries to 95 × 123 × 105 voxels—the dataset's largest edema extents. and hence ensured a global shape that was consistent for batching. Although this expansion added large amounts of healthy tissue for most patients, it still brought a more tumor-focused view and an 85 % size reduction.

MRI modalities (T1-CE, FLAIR) and segmentation maps were combined along the channel dimension, forming a 3 × 95 × 123 × 105 input per patient. Clinical features were processed by normalizing age and

dummy-encoding categorical variables.

Deep learning model

Model architecture

We employed a 3D Vision Transformer (ViT) [14] to encode MRI sequences into latent representations, which were fused with clinical features and processed through linear layers to predict LF risk. The encoder comprised twelve transformer layers (Fig. 1), using a latent size of 512, eight attention heads, and an MLP (multi-layer perceptron) dimension of 3072. These architectural parameters deviate from conventional ViT configurations [21] but were optimal for our use case. An ablation experiment showing this can be seen in supplementary Table C6.

Input images $x \in \mathrm{R}^{3\times95\times123\times105}$ were tokenized via non-overlapping patches using a single 3D convolution (16^3 kernel, stride 16) as in [14].

A learnable CLS (classification) token was appended to the initial patch embeddings and propagated through the transformer. This token acted as a *blank slate* for the input, forcing the model to encapsulate all vital information via self-attention. Ultimately, only the CLS token was used for prediction.

Following Darcet et al. [22], we appended four learnable register tokens to the end of the input sequence. These tokens help the transformer *register* redundant information to learn to ignore some details for more focused predictions. This approach also aids in isolating the network's attention around the target area, enhancing the explainability analysis.

The final CLS token state was concatenated with clinical features and passed through three LeakyReLU-activated [23] linear layers for dimensionality reduction before output.

All the learnable model parameters were initialized randomly without any transfer learning. Our model and workflow are visualized in Fig. 1.

Loss function

We modeled the LF risk as a scalar output of patient hazard ratios dependent on input features. Thus, we used *Cox Partial Log-Likelihood* [24] as the loss function to optimize our network. The equation can be found in supplementary material B.3.

Software details

We implemented the code in Python using *PyTorch* (version 2.2.0) as the deep learning framework. The model was based on the ViT from the *MONAI* library [25] (version 1.3.0), which was also used for data loading and augmentation. The training pipeline was created with *PyTorch Lightning* (version 1.8.6).

Training

All our experiments shared identical training configurations. The neural network weights were updated by batch-wise *gradient descent*. The *batch size* was ten, but we accumulated the gradients for four batches, increasing the effective size to 40 patients. Optimization-specific parameters can be seen in Appendix B.

We always trained as a five-fold cross-validation to inspect the variance in the prediction performance. The maximum number of epochs per fold was set to 100, which lasted around 20 min to complete. 6 GB of VRAM was sufficient to train the model.

Augmentations

To increase the robustness of models through more variability in the training data, we employed a set of random image-level intensity and

Table 1
Patient cohort demographics. Continuous features are given as medians and interquartile ranges (IQR) in parentheses. Categorical features are depicted by the number of counts in the dataset.

Training cohort				Test cohort					
	Overall $N = 253$	TUM[1] $N = 167$	USZ[2] $N = 86$	Overall $N = 99$	FD[3] $N = 5$	FFM[4] $N = 11$	FR[5] $N = 18$	HD[6] $N = 44$	KSA[7] $N = 21$
Age	62 (53,71)	62 (53,71)	62 (54,69)	61 (54,67)	63 (55,64)	57 (52,66)	58 (50,66)	61 (54,65)	63 (59,70)
KPS	80 (70,90)	80 (70,90)	90 (80,90)	90 (80,90)	80 (80,80)	90 (90,90)	90 (82,100)	80 (78,90)	90 (90,100)
Location									
Frontal	86	67	19	33	1	4	5	14	9
Temporal	32	18	14	7	2	0	1	2	2
Parietal	47	28	19	20	2	1	1	13	3
Occipital	27	12	15	12	0	2	3	5	2
Cerebellar	56	39	17	24	0	4	5	10	5
Other	5	3	2	3	0	0	3	0	0
Primary Diagnosis									
NSCLC	89	37	52	39	3	6	2	19	9
Melanoma	47	24	23	9	1	1	1	2	4
Breast	34	33	1	19	0	3	5	9	2
RCC	11	9	2	8	0	1	2	3	2
GI	26	26	0	11	0	0	4	5	1
Other	46	38	8	13	1	0	4	6	2
Residual Areas	66	66	0	21	1	2	1	11	6
Surgery to RT (d)	20 (5,29)	26 (20,34)	4 (3,5)	32 (22,44)	31 (28,32)	30 (24,40)	7 (6,8)	40 (31,50)	35 (25,44)
Concurrent CTX	15	8	7	3	0	2	0	1	0
Concurrent ITX	10	6	4	13	0	3	0	9	1
EQD2	43.75 (37.5, 43.75)	43.75 (43.75, 43.75)	37.5 (37.5, 37.5)	37.5 (34.7, 42.0)	37.5 (37.5, 40.0)	34.7 (28.9, 36.0)	37.5 (37.5, 42.3)	38.3 (34.7, 43.8)	40 (31.2, 40.0)
Total BM burden (ml)	11 (5,29)	11 (20,34)	12 (3,5)	13 (22,44)	41 (28,32)	17 (24,40)	14 (6,8)	9 (31,50)	14 (25,44)
Events	36	26	10	16	2	2	5	4	3

[1]: Technical University of Munich [2]: University Hospital of Zurich [3]: General Hospital Fulda [4]: University Hospital Frankfurt [5]: University Hospital Freiburg [6]: Heidelberg University Hospital [7]: Kantonsspital Aarau.

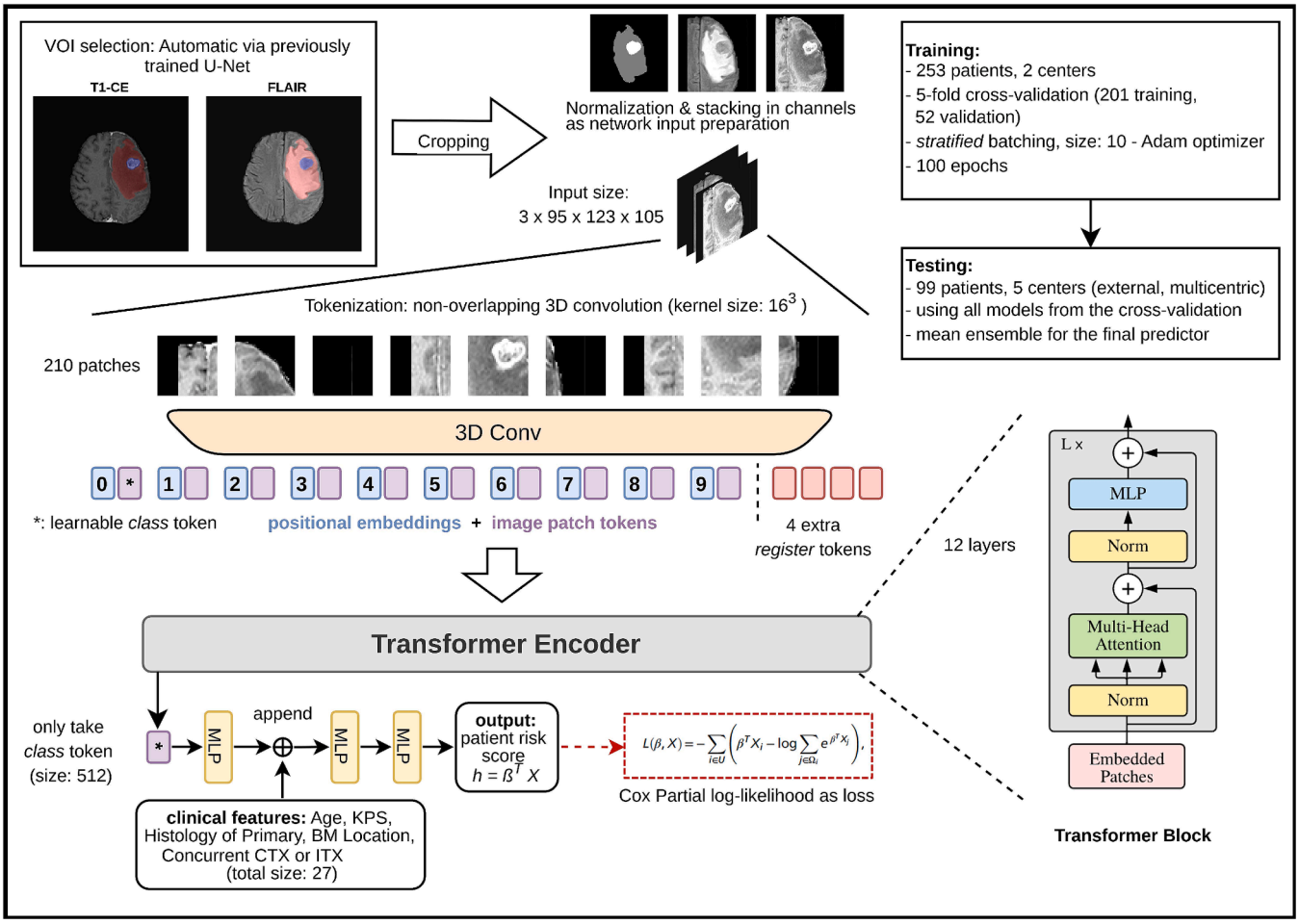


Fig. 1. Overview of our total workflow and our transformer architecture.

geometric augmentations. The total set of operations and their respective parameters can be viewed in supplementary Table B5.

Addressing LF Imbalance with batching

Patient distribution per batch is crucial since the loss value L_{Cox} becomes zero if all patients in a batch are censored, causing instability in the learning process. It is also resource-inefficient to run steps that do not contribute to the network updates. To address this, we used stratified batching, ensuring each batch included at least one patient (up to three to allow randomness) with LF as long as uncensored patients remained. If uncensored patients were oversampled, we started a new epoch.

Model ensemble

Combining multiple models into a single predictor to increase model performance is a common practice to reduce variance and exploit different focus points of diverse models [26]. Instead of picking the best, we performed ensembling by averaging the outputs of all folds for every test patient.

Evaluation

For evaluation, we used *Harrell's Concordance Index* (C-Index) [27], as typical in time-to-event analyses.

We compared our approach with various baseline methods. The predictive performance of two clinically established indices, Recursive Partitioning Analysis (RPA)[48] and Graded Prognostic Assessment (GPA)[49], as well as BM volume alone, was evaluated using univariate Cox analysis.

As deep learning baselines, we trained *DeepSurv* [28], a renowned time-to-event model for tabular data, with different sets of features: clinical only and clinical combined with image features extracted with a recently published medical *foundation model,* FMCIB [29].

Further, for end-to-end learning baselines, we used a 3D ResNet34 [30] and the *TransRP* [31] model, a hybrid network combining ResNet18 and ViT, which has been successful in predicting recurrence-free survival (RFS) in head-and-neck tumors. To ensure comparability, we applied the same input processing steps for the baseline models and modified their final predictor layers to match our approach (three linear layers, see Fig. 1). The method for fusing clinical features was also kept constant when applied. For all models, we report imaging-only results as well.

For patient stratification, the 33rd and 66th percentiles of the continuous risk ranks in the training cohort served as cutoffs, which were used to divide the test cohort into three groups based on predicted risk ranks. We employed Kaplan-Meier analysis to compare the survival rates among the groups. Both cutoff determination and stratification were performed on the final predictions after ensembling.

Model explainability

We conducted a qualitative analysis of the saliency maps from the networks to identify image areas relevant to predictions. In transformer architectures, saliency maps can be generated inherently from attention matrices of the layers as the network learns which patches are important. However, our model had twelve attention blocks with eight heads each, making it difficult to determine which is the *true explanation*.

Various methods have been explored [18] to incorporate more of the network's structure into the explanations. We applied the *beyond attention* method [32] that combines attention matrices with Layer-wise Relevance Propagation (LRP) [33] across all heads and layers to highlight the most relevant components in the image. The equations are provided in supplementary material B.4.

We considered relevance scores only for the CLS token against input tokens since the rest were discarded for the prediction at the end. Saliency maps were generated for one example model (with the highest test score), as it was challenging to backpropagate through the ensembled networks to create an aggregated explanation.

Results

Table 2 highlights the importance of our proposed methods. Each addition contributed to the total increase of C-Index from the baseline *vanilla* ViT (image only, no segmentation guidance, random batching) to our final model. We further show in the supplementary Table C7 that the improvement was model agnostic. Hence, when applicable, these features were set as default while benchmarking against other models.

As seen in Table 3, our ViT outperformed all the baseline models on average test set results (0.7336 C-Index). Compared to the image-only scenario, all the models benefited from including clinical features.

Among the clinical baselines, BM volume was a strong predictor in the test cohort with a 0.77 C-Index but only achieved 0.51 in the training cohort, showing high instability. There was no significant difference in BM volumes between the training and test cohorts ($p = 0.64$, Wilcoxon rank sum test).

Ensembling improved the model's performance in all cases up to the best total performance (0.7982 C-Index). This may be interpreted as the predominance of the best individual fold closer to the ensemble performance than the mean performance. There was no sign of a strong correlation between the ensemble predictions and the BM volumes (Spearman's rank coefficient $r = 0.2046$, $p = 0.042$).

In Table 4, we analyzed the performance of the model ensemble for each primary diagnosis subgroup. It was highly predictive for each primary cancer but was sensitive to ambiguity when multiple subtypes were gathered as *Other*.

When tested using expert-made contours instead of the U-Net segmentations, the performance increased to **0.7358** C-Index on average and to **0.7992** for the ensemble. In comparison with the expert segmentations, the U-Net contours had Dice coefficients of 0.8960 (edema) and 0.9266 (BM core) for the largest lesions in the test cohort.

Using the cutoffs from the training cohort, the test cohort was split into low-, medium-, and high-risk groups. Survival functions for each group are visualized in Fig. 2 with Kaplan-Meier analysis. Our model could stratify patients significantly ($p = 0.0001$, log-rank test).

Decision curve analysis indicated that our model provides a net benefit over treating all patients or none, and the clinical-only model within the threshold range from 0.1 to 0.5. Compared to the clinical-only model, it also shows an improved calibration (see Fig. 3).

Visualization of the relevant areas of the input MRI scans through gradient-fused attention maps [30] validated that our model focused well on the BM to predict the LF risk (Fig. 4). We investigated the overlap of BM volumes with the top 80th percentile of attention. Such a threshold was applied to remove regions with up to 10^3 smaller attention magnitudes.

On average, 89.5 % of the BM core and 78.4 % of GTV were covered by the attention area. Using the register tokens further increased the localization capabilities (93.5 % coverage of BM core and 84 % GTV). However, it should be noted that for both settings, attention was also observed outside of GTV areas and even outside of the brain structure. Specifically, 15 % of the total was laid over the GTV on average. We analyzed prediction quality by calculating individual C-Index scores per patient. There was no significant correlation between high C-Index and GTV coverage of attention (Spearman's rank coefficient $r = 0.25$, $p = 0.82$) or percentage of total attention laid over the GTV (Spearman's rank coefficient $r = 0.3$, $p = 0.88$).

Table 2
Mean ± standard deviation C-Index scores of our different Vision Transformer (ViT) settings on the held-out test set, illustrating the incremental performance gains from each methodological enhancement. Each row adds one feature to the previous configuration. The bolded value indicates the best-performing model variant.

Model	Added Feature	Test C-Index
Vanilla ViT	–	0.5912 ± 0.041
ViT	+ *Clinical Features*	0.6691 ± 0.031
ViT	+ *Segmentation-guided Image Cropping*	0.7139 ± 0.044
ViT	+ *Segmentations as Additional Input*	0.7267 ± 0.042
ViT	+ *Stratified Batching*	0.7311 ± 0.035
ViT	+ *Register Tokens*	**0.7336** ± 0.029

Table 3
Comparison against baseline models. Deep learning results are reported from cross-validation on the held-out test set and include the mean ± standard deviation, best-performing fold, and final performance of the ensemble predictor (averaged across folds). Bold values indicate the best performance in each column.

Model	Test C-Index		
Clinical (Cox analysis)			
Recursive Partitioning Analysis (RPA)[48]	0.39		
Graded Prognostic Assessment (GPA)[49]	0.44		
BM Volume	0.77		
Deep Learning	5-Folds	Best Fold	Ensemble
DeepSurv (clinical only)	0.6625 ± 0.023	0.7002	0.6681
DeepSurv (clinical + FMCIB)	0.5507 ± 0.016	0.5738	0.5549
ResNet34 (image only)	0.6501 ± 0.038	0.6845	0.6587
ResNet34 (image + clinical)	0.6879 ± 0.032	0.7105	0.7135
TransRP (image only)	0.6728 ± 0.021	0.6942	0.7022
TransRP (image + clinical)	0.7138 ± 0.049	**0.7915**	0.7401
ViT (image only)	0.6814 ± 0.038	0.7112	0.7144
ViT (image + clinical)	0.7311 ± 0.035	0.7784	0.7865
ViT + registers (image + clinical)	**0.7336** ± 0.029	0.7842	**0.7982**

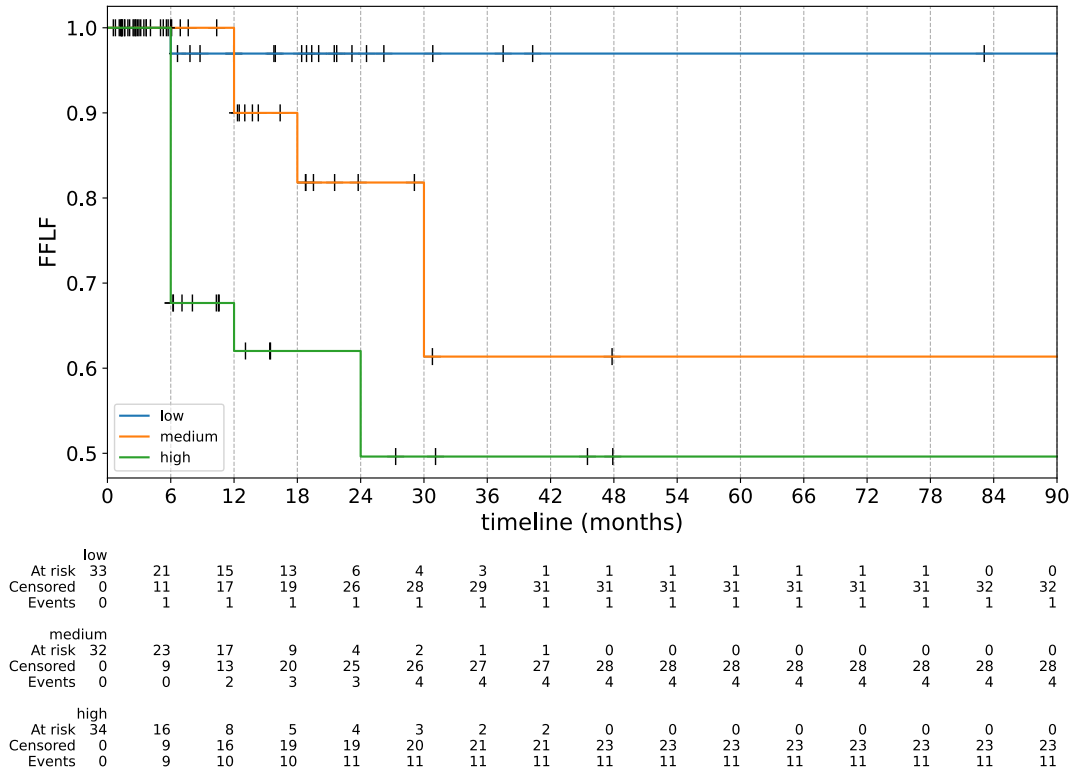
Table 4
Sensitivity of C-Index of the ensembled predictor on each primary diagnosis subgroup in the test cohort. No LF was observed among the melanoma patients, so C-Index could not be computed.

Primary Diagnosis	# of LFs	C-Index
NSCLC	5	0.8953
Melanoma	0	n.a.
Breast	3	0.8936
RCC	1	1.0
GI	4	0.88
Other	3	0.44

Discussion

We explored ViTs for predicting FFLF using pre-operative MRI scans of BM patients. Despite a relatively small dataset, our model benefited from network modifications and segmentation-guided input adaptations, achieving strong predictive performance (0.7982 C-Index). ViTs consistently outperformed clinical and CNN-based baselines, demonstrating their superiority.

However, larger models such as TransRP [31] and ViT-Basic [21] (with 118 M and 94.7 M parameters, respectively) struggled with data scarcity and showed lower test performance despite better training cohort learning. Our *medium-sized* ViT (57.3 M parameters) achieved a better balance.



low																
At risk	33	21	15	13	6	4	3	1	1	1	1	1	1	1	0	0
Censored	0	11	17	19	26	28	29	31	31	31	31	31	31	31	32	32
Events	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
medium																
At risk	32	23	17	9	4	2	1	1	0	0	0	0	0	0	0	0
Censored	0	9	13	20	25	26	27	27	28	28	28	28	28	28	28	28
Events	0	0	2	3	3	4	4	4	4	4	4	4	4	4	4	4
high																
At risk	34	16	8	5	4	3	2	2	0	0	0	0	0	0	0	0
Censored	0	9	16	19	19	20	21	21	23	23	23	23	23	23	23	23
Events	0	9	10	10	11	11	11	11	11	11	11	11	11	11	11	11

Fig. 2. Kaplan-Meier analysis of the final predictor results after ensembling. The 33rd and 66th percentiles of continuous hazard ratios in the training set were used as cut-offs.

Model robustness was validated via five-fold cross-validation, and ensembling further improved predictions by leveraging multiple models' strengths (Table 3). The ensembles were mostly governed by the highest-performing model. However, in edge cases, the versatility of multiple models helped align patient risks better and thus achieved a better overall ranking. Only TransRP ensembles showed performance declines, likely due to high variance among sub-models.

Our approach was also computationally efficient, requiring just 6 GB of VRAM and completing five-fold cross-validation in 100 min. Testing was rapid (0.5 s per patient), making it accessible for clinics with basic hardware.

Recent work [5] addressed the same problem with a clinically accustomed approach by using radiomics features and achieved a C-Index of 0.77 on the same test cohort. Radiomics requires domain expertise for feature selection, whereas our model operates easily and fast without intervention. Additionally, [5] showed low robustness to segmentation variances, dropping performance to a C-Index of 0.72 when using autosegmented contours. Our model maintained performance across segmentation differences and proved that it does not require hand-segmented contours. This could offer further ease in clinical workflow with a fully automatic pipeline.

Cox analysis with BM volume reached a 0.77 C-Index on the test set but only 0.51 in training, highlighting its limited generalizability. Saliency maps confirmed tumor-focused attention, aligning with expert observations. Using register tokens [22] further enhanced attention to relevant regions, improving GTV coverage (+7%) and predictive performance (+0.0117 C-Index).

Even though our model focused well on the BMs, a substantial amount of attention was also distributed in other regions. It should be noted that the generated maps serve as a means for interpreting the network's working mechanism and do not provide absolute reasoning for its decisions [18,32]. Complete explainability still remains a challenge for AI models in general due to the difficulty in interpreting the interactions across multiple layers and millions of parameters.

Clinically, patients with a high predicted risk of LF could benefit from tailored strategies such as SRT dose escalation or expanded clinical target volumes, both shown to improve local control [34]. Additional measures, including systemic agents crossing the blood–brain barrier and more frequent follow-ups, may further aid in early LF detection.

This work carries several limitations along with advances. The retrospective design led to suboptimal clinical data quality, necessitating prospective validation to confirm efficacy and reliability.

In clinical routine, diagnosing LF from radiation necrosis or pseudoprogression is challenging [35,36]. A significant portion of patients were labeled before the publication of the BM-RANO criteria [37]. Consequently, some cases may have been possibly misclassified, leading to noise in data, despite board-certified experts' diagnosis and, in part, histological validation. Our BM segmentation relied on a U-Net trained on the same dataset, potentially leading to idealized tumor contours. However, this U-Net was trained purely on the pre-therapeutic imaging data with no cases of prior radiation therapy. Therefore, the differentiation between metastases, pseudoprogression, or radiation necrosis has not biased the performance of the U-Net model.

In conclusion, our fully automated model outperformed existing radiomics-based approaches and demonstrated robust performance across a multicenter external test cohort, accommodating different MRI scanners and protocols. By tailoring treatment strategies to individual risk profiles, our model offers a promising step toward improved BM management.

Generative AI in scientific writing

During the preparation of this work the author(s) used ChatGPT-4o (OpenAI) for minor improvements in the language and readability of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

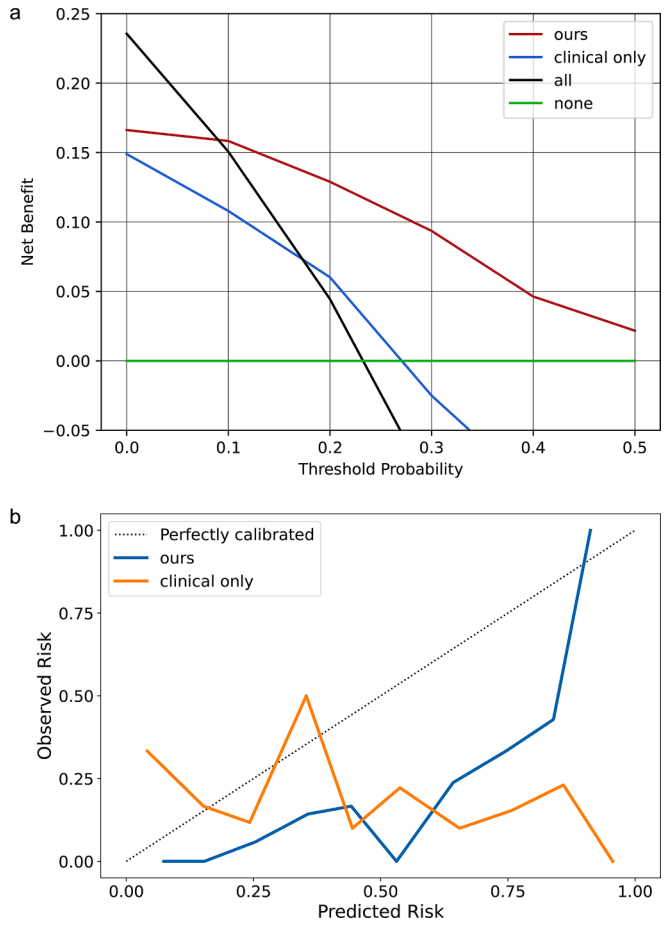


Fig. 3. Decision and calibration curves. DeepSurv model from Table 3 was taken as the *clinical-only* baseline. For the decision curve, the net benefit is calculated by subtracting the proportion of false-positive patients from the proportion of true-positive patients, weighted by the relative harm of a false-negative and false-positive result. The threshold probability was calculated for failure-free 24 months. Our model shows clinical benefit over the reference values of "treat all" and "treat none", and the clinical-only baseline as well, by having larger net benefit values. The predictions were transformed into event probabilities for the calibration curve. Compared to the clinical-only baseline, our model was closer to a perfectly calibrated predictor.

Code availability

The model may be accessed under https://github.com/RadOnc-AI/ViTs-BrainMet-Failure under the Commons Attribution-NonCommercial 4.0 International (CC BY NC 4.0) license.

CRediT authorship contribution statement

Ayhan Can Erdur: Writing – original draft, Methodology, Conceptualization, Writing – review & editing, Software, Data curation, Visualization, Formal analysis. **Daniel Scholz:** Software, Writing – original draft, Writing – review & editing, Methodology. **Q.Mai Nguyen:** Data curation, Writing – original draft, Writing – review & editing. **Josef A. Buchner:** Data curation. **Michael Mayinger:** Writing – review & editing, Data curation, Resources. **Sebastian M. Christ:** Writing – review & editing, Data curation, Resources. **Thomas B. Brunner:** Writing – review & editing, Data curation, Resources. **Andrea Wittig:** Writing – review & editing, Data curation, Resources. **Claus Zimmer:** Writing – review & editing, Data curation, Resources. **Bernhard Meyer:** Writing – review & editing, Data curation, Resources. **Matthias Guckenberger:** Writing – review & editing, Data curation, Resources. **Nicolaus Andratschke:** Writing – review & editing, Data curation, Resources. **Rami A. El Shafie:** Resources, Writing – review & editing, Data curation. **Jurgen Debus:** Resources, Writing – review & editing, Data curation. **Susanne Rogers:** Resources, Writing – review & editing, Data curation. **Oliver Riesterer:** Resources, Writing – review & editing, Data curation. **Katrin Schulze:** Resources, Writing – review & editing, Data curation. **Horst J. Feldmann:** Resources, Data curation. **Oliver Blanck:** Writing – review & editing, Data curation, Resources. **Constantinos Zamboglou:** Writing – review & editing, Data curation, Resources. **Angelika Bilger-Z.:** Writing – review & editing, Data curation, Resources. **Anca L. Grosu:** Writing – review & editing, Data curation, Resources. **Robert Wolff:** Data curation, Resources. **Kerstin A. Eitz:** Data curation, Resources. **Stephanie E. Combs:** Resources, Writing – review & editing, Data curation. **Denise Bernhardt:** Resources, Writing – review & editing, Data curation. **Benedikt Wiestler:** Supervision, Writing – review & editing, Funding acquisition. **Daniel Rueckert:** Resources, Supervision, Writing – review & editing, Funding acquisition. **Jan C. Peeken:** Supervision, Writing – original draft, Project administration, Conceptualization, Writing – review & editing, Resources, Formal analysis, Funding acquisition.

Ethics approval

Ethical approval was obtained at each institution (main approval at the Technical University of Munich: 119/19 S-SR). No informed consent for the research study was necessary as the retrospective analysis of patient records and data is generally allowed following Article 27 of the Bavarian Hospital Act (*Bayerisches Krankenhausgesetz*) from the law *Landeskrankenhausgesetz des Freistaates Bayern*. Informed consent for treatment was obtained from every patient.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project number 504,320,104 – PE 3303/1-1 (JCP), WI 4936/4-1 (BW), RU 1738/5-1 (DR)).

Declaration of competing interest

TBB: Honoraria: Merck, Takeda, Dalichi Sankyo

AW: Grants: EFRE, Siemens; Consulting fees: Gilead, Hologic Medicor GmbH; Honoraria: Accuray, Universitätsklinikum Leipzig AöR, Sanofi-Aventis GmbH; Travel support: DKFZ, DEGRO; Board: IKF GmbH (Krankenhaus Nordwest)

ClZ: Co-editor on the advisory board of "Clinical Neuroradiology", Leadership: President of the German society of Neuroradiology (DGNR)

BeM: Grants: BrainLab, Zeiss, Ulrich, Spineart; Royalities: Medacta, Spineart; Consulting fees and Honoraria: Medacta, Brainlab, Zeiss; Travel support: Brainlab, Medacta; Stock: Sonovum

MG: Grants: Varian/Siemens Healthineers, AstraZeneca, ViewRay Inc.; Honoraria: AstraZeneca; Leadership: ESTRO president elect, SAMO board member

NA: Grants: ViewRay Inc., AstraZeneca, SNF, SKL, University CRPP; Consulting Fees: ViewRay Inc., AstraZeneca; Honoraria: ViewRay Inc., AstraZeneca; Travel support: ViewRay Inc., AstraZeneca; Safety monitoring/advisory board: AstraZeneca, Equipment: ViewRay Inc.

RAES: Grants: Accuray; Consulting Fees: Novocure, Merck, AstraZeneca; Honoraria: Accuray, AstraZeneca, BMS, Novocure, Merck, Takeda; Travel support: Merck, Accuray, AstraZeneca; Safety monitoring/advisory board: Novocure, Merck; Stock: Novocure

JD: Grants: RaySearch Laboratories AB, Vision RT Limited, Merck Serono GmbH, Siemens Healthcare GmbH, PTW-Freiburg Dr. Pychlau GmbH, Accuray Incorporated; Leadership: CEO at HIT, Board of directors at University Hospital Heidelberg; Equipment: IntraOP

OB: Grants: STOPSTORM.eu; Leadership: Board member of the working groups for Stereotactic Radiotherapy of the German Radiation Oncology and Medical Physics Societies, Section Editor of

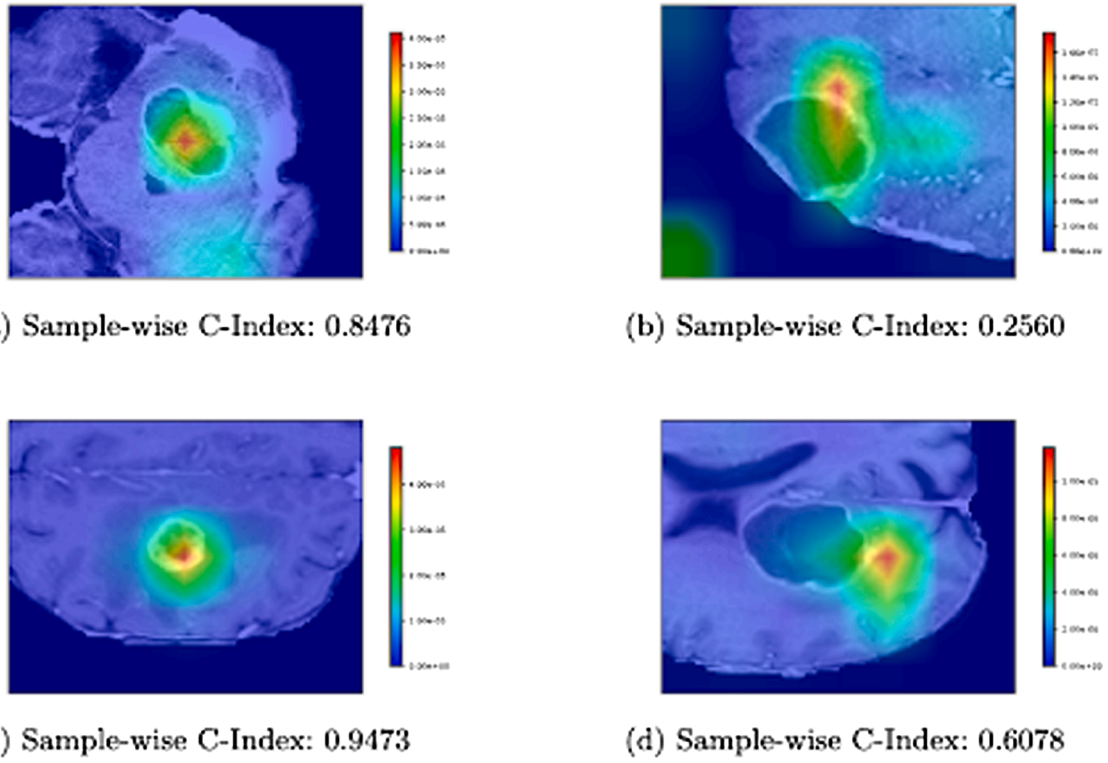


Fig. 4. Network relevance scores as heat maps overlaid on T1-CE scans for good predictions (left) and bad (right). Per-sample C-Index scores are used to estimate prediction quality.

"Strahlentherapie und Onkologie"

KF: Grants: Master of Disaster (Gyn Congress, Essen, Germany)

SEC: Grants, Consulting fees and Honoraria: Roche, AstraZeneca, Medac, Dr. Sennewald Medizintechnik, Elekta, Accuray, BMS, Brainlab, Daiichi Sankyo, Icotec AG, Carl Zeiss Meditec AG, HMG Systems Engineering, Janssen; Safety monitoring/advisory board: CureVac DSMB Member; Leadership: NOA Board Member, DEGRO Board Member

DR: Grants: DFG, ERC, EPSRC, BMBF, Alexander von Humboldt Stiftung; Consulting fees: ERC

BW: Grants: DFG, NIH, Deutsche Krebshilfe, BMWi; Consulting fees and Stock: Need; Honoraria: Philips, Novartis

JP: Honoraria: AstraZeneca, Support for current manuscript: German Research Foundation

The remaining authors have no potential conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2025.111031.

References

[1] Sacks P, Rahman M. Epidemiology of brain metastases. Neurosurgery Clin 2020;31:481–8.
[2] Lin X, DeAngelis LM. Treatment of brain metastases. J Clin Oncol 2015;33:3475–84.
[3] Vogelbaum MA, Brown PD, Messersmith H, Brastianos PK, Burri S, Cahill D. et al. Treatment for brain metastases: ASCO-SNO-ASTRO guideline, Oxford University Press US, 2022.
[4] Minniti G, Niyazi M, Andratschke N, Guckenberger M, Palmer JD, Shih HA, et al. Brown and others, "current status and recent advances in resection cavity irradiation of brain metastases,". Radiat. Oncol. 2021;16:1–14.
[5] Radiomics-based prediction of local control in patients with brain metastases following postoperative stereotactic radiotherapy, Neuro-oncology, p. noae098, 2024.
[6] Du P, Liu X, Shen L, Wu X, Chen J, Chen L, et al. Prediction of treatment response in patients with brain metastasis receiving stereotactic radiosurgery based on pre-treatment multimodal MRI radiomics and clinical risk factors: a machine learning model. Front Oncol 2023;13:1114194.
[7] Mulford K, Chen C, Dusenbery K, Yuan J, Hunt MA, Chen CC, et al. A radiomics-based model for predicting local control of resected brain metastases receiving adjuvant SRS. Clin Trans Radiat Oncol 2021;29:27–32.
[8] Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I, et al. Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery. Neuro Oncol 2020;22:797–805.
[9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[10] Jalalifar SA, Soliman H, Sahgal A, Sadeghi-Naini A. Predicting the outcome of radiotherapy in brain metastasis by integrating the clinical and MRI-based deep learning features. Med Phys 2022;49:7167–78.
[11] Cha YJ, Jang WI, Kim M-S, Yoo HJ, Paik EK, Jeong HK, et al. Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks. Anticancer Res 2018;38:5437–45.
[12] Buzea CG, Buga R, Paun M-A, Albu M, Iancu DT, Dobrovat B, et al. AI evaluation of imaging factors in the evolution of stage-treated metastases using Gamma Knife. Diagnostics 2023;13:2853.
[13] Zhao J, Vaios E, Wang Y, Yang Z, Cui Y, Reitman ZJ, et al. Dose-incorporated deep ensemble learning for improving brain metastasis stereotactic radiosurgery outcome prediction. Int J Radiat Oncol* Biol* Phys 2024;120:603–13.
[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[15] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. in International conference on machine learning, 2021.
[16] Castelvecchi D. Can we open the black box of AI? Nat News 2016;538:20.
[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017;30.
[18] Kashefi R, Barekatain L, Sabokrou M, Aghaeipoor F. Explainability of vision transformers a comprehensive review and new perspectives. arXiv preprint arXiv:231106786 2023.
[19] Buchner JA, Kofler F, Etzel L, Mayinger M, Christ SM, Brunner TB, et al. Development and external validation of an MRI-based neural network for brain metastasis segmentation in the AURORA multicenter study. Radiother. Oncol. 2023;178:109425.
[20] Kofler F, Berger C, Waldmannstetter D, Lipkova J, Ezhov I, Tetteh G. Brats toolkit: translating brats brain tumor segmentation algorithms into clinical and scientific practice. Front Neurosci 2020;125.
[21] Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L. How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270, 2021.
[22] Darcet T, Oquab M, Mairal J, Bojanowski P. Vision transformers need registers. arXiv preprint arXiv:2309.16588, 2023.

[23] Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853, 2015.
[24] Cox DR. Partial likelihood. Biometrika 1975;62:269–76.
[25] T. M. O. N. A. I. Consortium, Project MONAI, Zenodo, 2020.
[26] Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. J King Saud Univ - Comput Informat Sci 2023;35:757–74.
[27] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. J Am Med Assoc 1982;247:2543–6.
[28] Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Method 2018;18:1–12.
[29] Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa TL, et al. Birkbak and others, "Foundation model for cancer imaging biomarkers. Nat Mach Intell 2024;6:354–67.
[30] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[31] Ma B, Guo J, van Dijk LV, van Ooijen P, Both S, Sijtsema NM. TransRP: Transformer-based PET/CT feature extraction incorporating clinical data for recurrence-free survival prediction in oropharyngeal cancer. in Medical Imaging and Deep Learning, 2023.
[32] Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.
[33] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10:e0130140.
[34] Choi CYH, Chang SD, Gibbs IC, Adler JR, Harsh IV GR, Lieberson RE, et al. "Stereotactic radiosurgery of the postoperative resection cavity for brain metastases: prospective evaluation of target margin on tumor control. Int J Radiat Oncol* Biol* Phys 2012;84:336–42.
[35] Bernhardt D, König L, Grosu A-L, Rieken S, Krieg SM, Wick W, et al. DEGRO practical guideline for central nervous system radiation necrosis part 2: treatment. Strahlenther Onkol 2022;198:971–80.
[36] Bernhardt D, König L, Grosu A, Wiestler B, Rieken S, Wick W, et al. DEGRO practical guideline for central nervous system radiation necrosis part 1: classification and a multistep approach for diagnosis. Strahlenther Onkol 2022;198:873–83.
[37] Lin NU, Lee EQ, Aoyama H, Barani IJ, Barboriak DP, Baumert BG, et al. Response assessment criteria for brain metastases: proposal from the RANO group. Lancet Oncol 2015;16:e270–8.