Original Article | Open Access

Artificial Intelligence Surgery

Makolli *et al. Art Int Surg.* 2026;6:665-81 **DOI:**10.20517/ais.2026.26

Artificial intelligence-based intraoperative microscope image analysis for differentiating glioblastoma and brain metastases

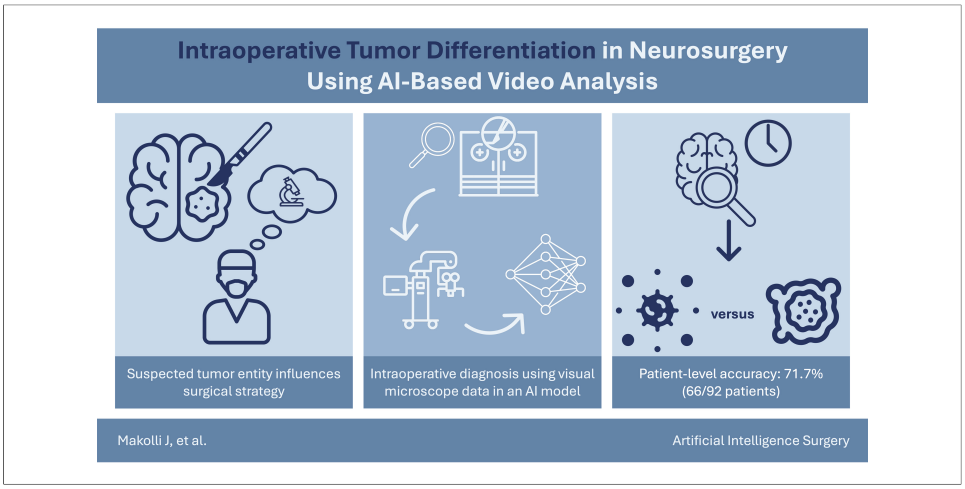
Jessica Makolli[1], Neal Bacher[2], Stefan Saur[2], Nathalia Spier[3], Ammar Mallouhi[4], Christian Dorfer[1,5]

Keywords:
Artificial intelligence, multimodal learning, intraoperative diagnosis, glioblastoma, brain metastasis, neurosurgical decision-making

Citation: Makolli J, Bacher N, Saur S, Spier N, Mallouhi A, Dorfer C. Artificial intelligence-based intraoperative microscope image analysis for differentiating glioblastoma and brain metastases. *Art Int Surg.* 2026;6:665-81.
https://dx.doi.org/10.20517/ais.2026.26

Received: 20 Apr 2026
First Decision: 20 Aug 2026
Revised: 30 Sep 2026
Accepted: 29 Oct 2026
Published: 17 Nov 2026

Academic Editor:
Andrew Gumbs
Copy Editor:
Tong Wang
Production Editor:
Tong Wang



Abstract

Aim: Intraoperative differentiation between glioblastoma and brain metastases may influence surgical strategy. This study aimed to develop and internally validate an artificial intelligence (AI)-based model combining intraoperative microscopic images with clinical metadata for differentiation of glioblastoma and brain metastases.

Methods: This retrospective single-center proof-of-concept study included 92 patients (46 glioblastomas, 46 metastases). A dual-branch neural network combining DINOv2 image features and clinical metadata was evaluated using patient-stratified 10-fold cross-validation. Patient-level classifications were aggregated by majority vote in a separate evaluation. Comparisons included a metadata-only model, an image-only ablation model and an information-restricted assessment by 13 neurosurgeons.

Results: Pooled analysis of 6,984 out-of-fold frame predictions yielded balanced accuracy

[1]Department of Neurosurgery, Medical University of Vienna, Vienna 1090, Austria.
[2]Carl Zeiss Meditec AG, Oberkochen 73447, Germany.
[3]Carl Zeiss AG, Oberkochen 73447, Germany.
[4]Department of Biomedical Imaging and Image-guided Therapy, Division of Neuroradiology, Medical University of Vienna, Vienna 1090, Austria.
[5]Department of Neurosurgery, University Hospital St. Pölten, Karl Landsteiner University for Health Sciences, St. Pölten 3100, Austria.

Correspondence to: Prof. Dorfer Christian, Department of Neurosurgery, Medical University of Vienna, Vienna 1090, Austria. E-mail: christian.dorfer@stpoelten.lknoe.at

of 68.2%, with an area under the curve (AUC) of 0.726 (95%CI: 0.639-0.806). Patient-level accuracy was 71.7% (66/92), with an AUC of 0.791 (95%CI: 0.694-0.879). The metadata-only model achieved 70.7% accuracy (65/92) and an AUC of 0.779 (95%CI: 0.678-0.869). Classifications differed in 19/92 patients (20.7%). In the matched ablation analysis, mean frame-level accuracy was higher for the multimodal than image-only model (67.9% *vs.* 61.2%). Mean neurosurgeon accuracy was 52.8%, compared with 72.1% for a separately trained image-only hold-out model evaluated on the same selected frames.

Conclusion: Multimodal analysis of selected intraoperative microscope images was technically feasible, but adding images to clinical metadata did not demonstrate an incremental benefit in patient-level classification accuracy. Prospective multicenter studies with external validation and assessment in continuous intraoperative workflows are required before clinical utility can be established.

INTRODUCTION

Advanced preoperative imaging techniques, including magnetic resonance imaging (MRI) and positron emission tomography-computed tomography (PET-CT), can provide valuable information to distinguish glioblastoma (GBM) from brain metastases. However, several recent studies have demonstrated that definitive discrimination between these tumor types is not always possible without direct histopathological analysis and that preoperative diagnosis is sometimes uncertain[1,2]. While circumferential, perilesional resection techniques and extended resection strategies have been investigated in GBM, surgery for brain metastases may favor en bloc resection with circumferential dissection along the tumor-brain interface when anatomically feasible[3-5]. Therefore, the ability to differentiate between tumor types during surgery may affect surgical strategies, highlighting the importance of reliable intraoperative diagnostic tools[6].

Histopathological frozen section analysis remains the standard approach for intraoperative diagnosis. However, this method is inherently labor-intensive, frequently requiring up to 60 min for processing and interpretation, and relies on the availability of specialized neuropathologists[7,8]. The diagnostic accuracy of frozen section analysis may be influenced by factors such as tissue sampling, prior treatment-related changes, and variability in histopathological interpretation[9,10]. Furthermore, certain central nervous system (CNS) tumor entities present particular diagnostic challenges, including the differentiation of gliomas from reactive changes[10]. Additionally, these limitations, together with tissue consumption and processing artifacts, highlight the potential value of complementary rapid diagnostic approaches, including methods that enable remote neuropathological assessment[11].

To form a preliminary opinion on a tumor's histological entity during resection, surgeons traditionally rely on macroscopic tumor characteristics, such as color, vascularization patterns, tissue consistency and adherence to surrounding structures[12,13]. Despite advanced preoperative imaging and surgical experience, the assessment of such intraoperative characteristics remains largely subjective, highlighting the necessity for more objective and reproducible tools that can help support intraoperative tumor diagnosis.

Recent advancements in computational data-driven image processing have made it possible to extract a wide range of complex visual information from medical images that can be seamlessly used in many areas of medicine, including pathology, radiology, and surgery[14]. In automated tumor classification, such techniques have demonstrated high accuracy by analyzing radiological images, pathology slides, and patient data[14]. The integration of such techniques into the intraoperative environment provides the potential to use both quantitative and objective image-based analysis to supplement the surgeon's subjective visual assessment during tumor removal[15]. However, model performance may vary across patient populations due to differences in demographic and clinical characteristics, highlighting the importance of assessing potential

disparities and algorithmic fairness during clinical validation. Whether such disparities exist in intraoperative video-based differentiation of glioblastoma and brain metastases remains unclear.

Recently, vision transformer (ViT) architectures have demonstrated strong performance for interpreting complex surgical video data[16]. Unlike traditional convolutional neural networks (CNNs), which build spatial context hierarchically through local convolutional filters and require many stacked layers to achieve a large receptive field, ViTs leverage self-attention mechanisms to model long-range spatial dependencies from the earliest layers. This inductive bias towards global context can be advantageous for capturing spatial relationships and contextual information across different regions of medical images[17]. In conjunction with specific clinical parameters, these artificial intelligence (AI) models may provide objective diagnostic information during surgery. Furthermore, automated tumor classification using such techniques may bridge the critical time gap in surgery, supporting intraoperative decision-making until surgeons receive results from frozen section analysis[15].

For these reasons, the primary purpose of this study was to develop and internally validate an AI-based model capable of differentiating glioblastoma from brain metastases, using high-resolution intraoperative surgical microscope video data in combination with relevant clinical information in a retrospective single-center proof-of-concept setting.

METHODS

Data collection

This was a retrospective, single-center proof-of-concept study with internal validation conducted at the Medical University of Vienna. This retrospective study was not registered in a public study registry. A total of 532 patients with glioblastoma or brain metastasis treated between January 2020 and December 2023 were initially assessed for eligibility [Figure 1]. Of these, 244 patients had brain metastases and 288 had glioblastoma. Patients were included if they (1) were aged 18 years or older, (2) had a definitive postoperative histopathological diagnosis of either glioblastoma or brain metastasis, and (3) had available high-resolution intraoperative video recordings and preoperative MRI scans. Patients undergoing stereotactic biopsy and patients without available intraoperative video data or preoperative MRI were excluded, leaving 120 patients for video quality assessment. Patients were excluded at this stage if video data were of insufficient quality due to excessive hemorrhage, smoke artifacts, or persistent defocusing. Of these 120 patients, 28 were excluded due to insufficient video quality, resulting in a final study cohort of 92 patients who underwent resection, comprising 46 patients with glioblastoma and 46 patients with brain metastases. All 46 glioblastomas were isocitrate dehydrogenase (IDH)-wildtype and classified according to the *WHO Classification of Tumours of the Central Nervous System*, fifth edition (2021). No formal sample size calculation was performed. The study size was determined by the number of eligible patients with available video and clinical data during the predefined study period. Given the proof-of-concept design, all patients meeting the predefined eligibility and video quality criteria were included. The resulting cohort of 92 patients allowed an initial assessment of model feasibility and performance but was insufficient to establish generalizability or definitive clinical performance. Video data were sourced from the institutional surgical video database. Clinical and radiological metadata, including age, sex, location of tumor, patterns of contrast enhancement, and history of seizures before surgery, were obtained from institutional electronic health records and preoperative MRI scans. Preoperative radiological diagnoses were retrospectively extracted from available neuroradiological reports. Concordance with the final histopathological tumor category was defined as suspicion of a glial tumor in glioblastoma cases or suspicion of metastatic disease in brain metastasis cases. No predictor data were missing in the final cohort. Therefore, no imputation was performed. Patients and members of the public were not involved in the design, conduct, reporting, or interpretation of this study. The study was approved by the Ethics Committee of the Medical University of Vienna (Protocol No. 1250/2024) as a

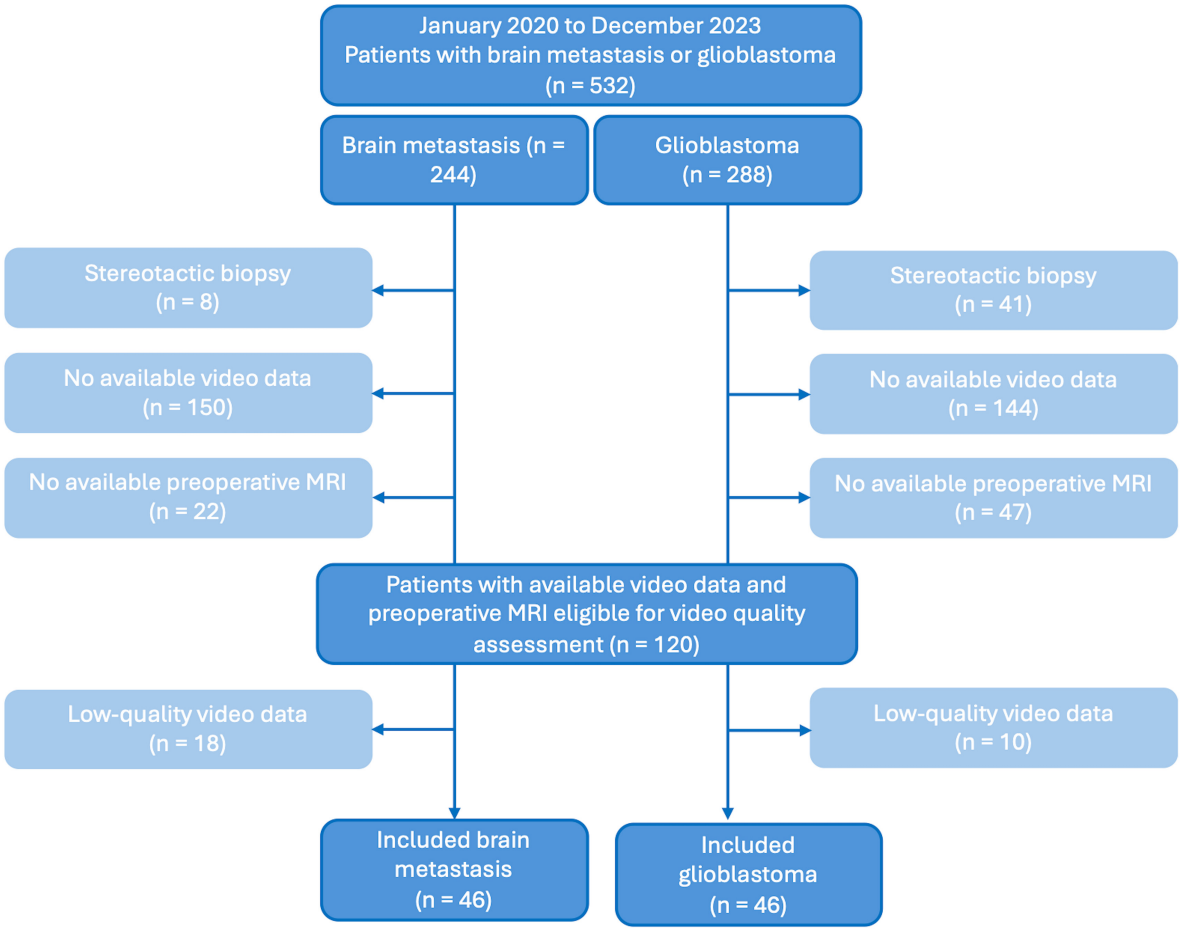


Figure 1. Flowchart of patient selection. A total of 532 patients with glioblastoma or brain metastasis treated between January 2020 and December 2023 were assessed for eligibility. After application of the exclusion criteria and assessment of video quality, 92 patients (46 with glioblastoma and 46 with brain metastasis) were included in the final study cohort.

retrospective analysis of previously acquired intraoperative video and clinical data. Written informed consent was obtained from all patients for the recording of intraoperative videos and their use for scientific research purposes.

Frame selection and annotation

Frames were extracted from the intraoperative video recordings at a sampling rate of 1 frame per second (1 fps) using an annotation tool provided by ZEISS. Each extracted frame was subsequently assessed individually for selection and annotation. To produce a high-quality data set, a senior neurosurgeon evaluated each video with the aim of selecting approximately 100 informative frames per case whenever image quality permitted. The number of retained frames varied between cases because only frames meeting the predefined image-quality criteria were included. Frame selection was performed as the first step of image annotation, with the neurosurgeon blinded to the final histopathological diagnosis. The selected frames demonstrated informative visual features of the tumors, such as specific tissue textures and vascularization patterns. Frames failing to demonstrate the tumor-brain interface without significant distortion were excluded. Images that were distorted by excessive bleeding, blurriness, overexposure, or smoke from electrocautery were also removed from the dataset. Tumor type labels (glioblastoma or metastasis) were assigned based on the final histopathologic report. Following this quality-based selection, 6,984 frames from 92 patients were included in the final out-of-fold (OOF) frame-level evaluation. The separate patient-level

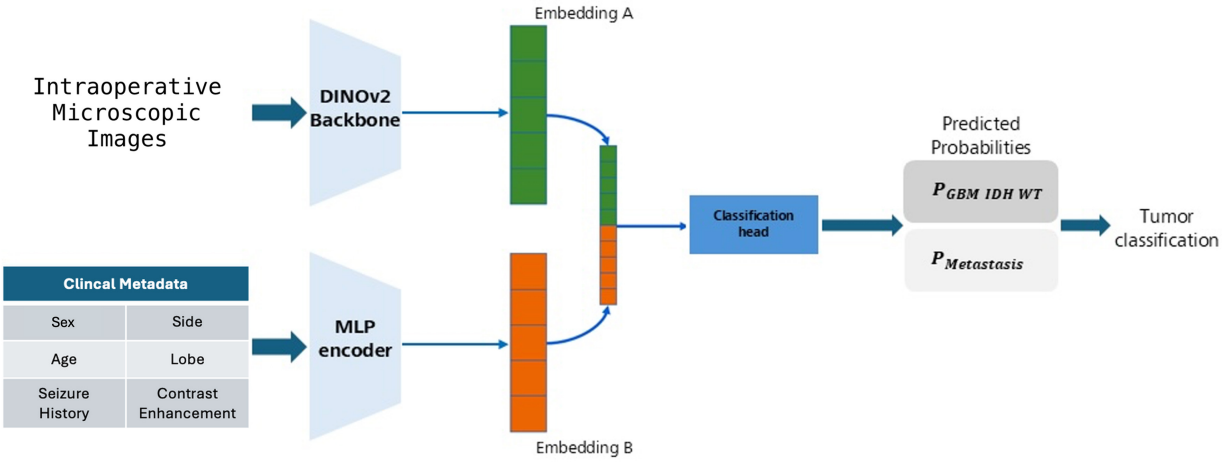


Figure 2. AI-based classification pipeline combining intraoperative microscopic images and clinical parameters to differentiate GBM from brain metastases. MLP: Multilayer perceptron; AI: artificial intelligence; GBM: glioblastoma.

evaluation used the annotated frames available for video-level classification without applying the same additional frame-level quality filtering. Therefore, the two analyses were not based on identical frame sets. The quality-selected frames were used for frame-level model development and evaluation, while a separate manually selected subset was used for the comparative reader assessment. Thus, the dataset represented a manually enriched selection of informative intraoperative frames rather than unselected continuous operative video.

AI model development

The AI model was developed with a dual-branch architecture that was specifically designed to integrate visual and clinical data. The first branch uses the vision transformer network DINOv2 to develop high-dimensional semantic representations (Embedding A) from the intraoperative images. The second branch uses a multilayer perceptron (MLP) encoder to develop clinical feature vectors (Embedding B) from the structured clinical data, including sex, age, location of tumor (lobe/side), contrast enhancement, and seizure history. These predictors were selected from routinely available preoperative clinical and radiological data. Both Embeddings A and B were combined via a late fusion strategy. The feature embedding outputs from both branches are merged and passed through a fully connected layer for binary classification to differentiate between glioblastomas and metastases. This integrated approach allows the AI to combine image-based insights with clinical factors, such as tumor location, contrast enhancement, and seizure history for multimodal classification [Figure 2].

Model architecture and training details

The architecture of the model employs a frozen, pre-trained DINOv2 ViT-S/14 (patch size 14) backbone as a robust feature extractor of visual information. DINOv2 was chosen due to its superior ability to extract high-level semantic features from complex visual textures. The selected frames were resized to 224 × 224 pixels using bicubic interpolation, and normalized using ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Clinical metadata is processed separately through a parallel MLP branch, with three hidden layers of 32 neurons. The output of the clinical branch is concatenated with the visual feature vector. Both are then fed into a fusion MLP, which contains six layers with a hidden dimension of 128 neurons. To prevent overfitting, a dropout rate of 0.2 was applied across all fully connected layers.

Due to the frozen nature of the feature extractor, the number of trainable parameters was low, and therefore the training process was very rapid. Each training session took approximately 5 min on an NVIDIA RTX 3090 graphics processing unit (GPU). All models were trained for 120 epochs using cross-entropy as the

primary loss function and a batch size of 8. AdamW was used with a learning rate of 1×10^{-4} and weight decay of 1×10^{-3}. Data augmentation included horizontal and vertical flips, rotation (±20°), Gaussian noise, coarse dropout (3-10 holes, 32-64 pixels), brightness/contrast shifts (±30%), and blur. Hyperparameters were tuned by manual grid search over MLP depth, hidden-layer width (64, 128, or 256), and dropout (0.2 or 0.4), with the DINOv2 backbone remaining frozen. Given the limited cohort size, the manual grid search was evaluated using the same patient-stratified cross-validation folds subsequently used to report final model performance. The final configuration used a six-layer MLP with a hidden dimension of 128 and dropout of 0.2. No separate validation set or early stopping was used, and the model obtained after 120 epochs was retained. Classification was based on softmax probabilities using a threshold of 0.5. To address the limited data volume available and to obtain an internal estimate of model performance, a patient-stratified 10-fold cross-validation strategy was used to divide the dataset into different training and test sets. In order to prevent any potential data leakage in which information from a specific patient's training sample is used to predict the outcome of a frame-based feature on the same patient, the stratification of the data was performed at the patient level. Therefore, 10 individual models were created to verify consistent performance. Models were implemented in PyTorch 2.2.x with Python 3.10.14, CUDA 11.8, and cuDNN 8.7.0. Source code and trained model weights are not publicly available.

Patient-level prediction and ablation study design

Patient-level performance was evaluated separately from the quality-filtered frame-level analysis. For each patient, the classifications of the annotated frames available for video classification were aggregated by majority voting to obtain a single patient-level prediction. This analysis was restricted to the same 92 patients included in the final frame-level cohort. In addition, an ablation analysis was performed by removing the clinical metadata branch and comparing the resulting image-only model with the multimodal model. For direct comparison, the ablation analysis was restricted to identical frames for which predictions from both model variants were available within the corresponding patient-stratified cross-validation folds. This resulted in a matched subset of 6,969 frames from 90 patients. The reduction from the 6,984 frames and 92 patients included in the primary multimodal frame-level analysis occurred exclusively in fold 1. For two patients in this fold, the frame sets evaluated by the multimodal and image-only models did not overlap, precluding a matched frame-by-frame comparison and resulting in the exclusion of 15 multimodal frames from the matched analysis. In folds 2-10, the evaluated frame sets were identical between both model variants. The discrepancy in fold 1 resulted from differences in the quality-based frame selection between the two model evaluations. Differences in accuracy and balanced accuracy between the multimodal and image-only models were calculated for each matched fold and assessed using paired *t*-tests.

Metadata-only comparator

To quantify the discriminative information contained in the clinical metadata alone, a metadata-only ridge-penalized logistic regression was evaluated using the same 92-patient cohort. Predictors comprised age, sex, tumor location (lobe and side), contrast enhancement, and seizure history. Ridge penalization was selected because conventional maximum-likelihood logistic regression showed sparse-data instability and failed to converge. The predefined patient-level 10-fold partition was used for outer cross-validation, with the ridge penalty parameter (lambda) selected by 5-fold cross-validation within each training fold. A probability threshold of 0.5 was applied. Sensitivity analyses included nested leave-one-out cross-validation and location-only models using lobe alone or lobe plus side.

Establishment of a visual diagnostic baseline

To establish a human visual baseline, 13 neurosurgeons (five residents, four junior board-certified neurosurgeons, and four experienced board-certified neurosurgeons) independently reviewed 33 intraoperative frames derived from 12 cases and classified each frame as either glioblastoma or brain

metastasis. The same 33 frames were presented to all readers. The evaluation set comprised 16 glioblastoma and 17 brain metastasis frames. Frames were manually selected from the previously annotated image set to represent favorable intraoperative visualization conditions, including minimal blood contamination, good illumination, and, where available, adjacent normal cortex. This selection was intended to provide a best-case scenario for visual differentiation of the two tumor types. Readers were provided with the intraoperative frames only and received no clinical, radiological, or histopathological information. Reader performance was calculated as the proportion of correctly classified frames and was summarized descriptively overall and according to level of clinical experience.

For comparison with human visual assessment, a separate image-only hold-out AI model was trained without clinical metadata and without any interaction with the 12 cases included in the reader evaluation. The model was evaluated at the frame level on the same 33 frames across five independently trained runs. AI and neurosurgeon performance were compared descriptively; no inferential statistical comparison between the AI and human readers or between reader experience groups was performed.

Statistical analysis

Model performance was assessed using accuracy, balanced accuracy, sensitivity, specificity, precision, F1 score, and area under the curve (AUC), as appropriate. Frame-level overall performance was calculated from pooled OOF predictions across the patient-stratified 10-fold cross-validation. Patient-level performance was evaluated separately by majority-vote aggregation of the annotated frames available for the video-level evaluation, restricted to the same 92-patient cohort. In the event of a tie, the respective patient would have been excluded from the patient-level analysis. However, no ties occurred, and no patients were excluded for this reason. For the matched ablation analysis, fold-specific differences between the multimodal and image-only models were assessed using paired *t*-tests and are reported as mean paired differences with 95% confidence intervals. Neurosurgeon and hold-out AI performance in the reader assessment were summarized descriptively, without inferential statistical comparison. All tests were two-sided, and $P < 0.05$ was considered statistically significant. The metadata-only comparator was evaluated using accuracy and AUC, and its patient-level classifications were compared with those of the multimodal model using an exact McNemar test. Frame-level AUC was calculated from continuous prediction scores, while sensitivity, specificity, and precision were based on binary predictions. Their 95%CIs were estimated using patient-level cluster bootstrapping (10,000 iterations) to account for within-patient frame correlation. For the multimodal patient-level ROC-AUC, continuous frame-level prediction scores were averaged within each patient to obtain a single patient-level score. Patient-level AUCs were calculated from continuous prediction scores for both the multimodal and metadata-only models, with 95%CIs estimated using stratified patient-level bootstrapping (10,000 iterations).

RESULTS

Patient characteristics

The final study cohort comprised 92 patients, consisting of 46 glioblastomas and 46 brain metastases. Patients with glioblastoma had a median age of 65.6 years [interquartile range (IQR), 57.9-72.2], and 20 (43.5%) were female, whereas patients with brain metastases had a median age of 61.7 years (IQR, 57.4-67.3), and 26 (56.5%) were female. Detailed demographic and tumor characteristics stratified by tumor type are presented in Table 1. Preoperative radiological diagnoses were available for 77 of 92 patients (83.7%). The suspected diagnosis was concordant with the final tumor category in 70 of 77 cases (90.9%), including 31 of 36 glioblastomas (86.1%) and 39 of 41 brain metastases (95.1%). Among glioblastoma patients, glioblastoma was explicitly suspected in 20 of 36 cases (55.6%). Among patients with glioblastoma, the O^6-methylguanine-DNA methyltransferase (MGMT) promoter was methylated in 27 (58.7%) and unmethylated in 19 (41.3%). The metastasis group exhibited a high degree of primary tumor diversity, mostly originating from the lung (*n*

Table 1. Baseline characteristics of the study cohort stratified by tumor type

	Glioblastoma (*n* = 46)	Brain metastasis (*n* = 46)
Age, years, median (IQR)	65.6 (57.9-72.2)	61.7 (57.4-67.3)
Female sex, *n* (%)	20 (43.5)	26 (56.5)
Male sex, *n* (%)	26 (56.5)	20 (43.5)
Tumor location, *n* (%)		
Frontal	13 (28.3)	11 (23.9)
Temporal	14 (30.4)	8 (17.4)
Parietal	3 (6.5)	10 (21.7)
Occipital	4 (8.7)	4 (8.7)
Cerebellar	1 (2.2)	11 (23.9)
Central	11 (23.9)	2 (4.3)
Present contrast enhancement, *n* (%)	43 (93.5)	46 (100.0)
Preoperative seizure history, *n* (%)	12 (26.1)	1 (2.2)
Preoperative radiological diagnosis, *n* (%)		
Available	36 (78.3)	41 (89.1)
Concordant with tumor category, *n/N* (%)[1]	31/36 (86.1)	39/41 (95.1)
Explicit suspicion of glioblastoma, *n/N* (%)[2]	20/36 (55.6)	-

Values are presented as *n* (%) unless otherwise specified. Contrast enhancement and preoperative seizure history were categorized as present or absent. [1]Radiological concordance was defined as suspected glial tumor in glioblastoma cases or suspected metastasis in metastatic cases; percentages refer to patients with available radiological assessments. [2]Explicit GBM suspicion required glioblastoma to be specifically stated rather than a broader glioma diagnosis. GBM: Glioblastoma; IQR: interquartile range.

= 23, 50.0%), followed by breast and colon (*n* = 5 each, 10.9%), melanoma (*n* = 3, 6.5%), kidney and ovary (*n* = 2 each, 4.3%) and single cases of pharyngeal, laryngeal, pancreatic, cervical, bladder and neuroendocrine tumors.

Quantitative performance of the AI model

The developed multimodal AI system was capable of differentiating between the two tumor types at the frame level. Frame-level performance was evaluated using patient-stratified 10-fold cross-validation, with all frames from an individual patient restricted to a single fold. The quantitative performance metrics for the ten cross-validation folds at the frame level are presented in Table 2. Mean frame-level balanced accuracy across the ten folds was 67.1%, ranging from 51.6% to 83.0%, while mean frame-level accuracy was 67.7%, ranging from 52.1% to 83.0%.

To obtain an overall frame-level performance estimate across the complete cross-validation cohort, the held-out test predictions from all ten folds were pooled. This resulted in 6,984 OOF frame-level predictions from 92 patients, with each frame evaluated only when its respective patient was part of the held-out test fold. Pooled frame-level OOF accuracy was 68.2%, with a balanced accuracy of 68.2%, sensitivity of 67.8% (95%CI: 57.5%-77.8%), specificity of 68.6% (95%CI: 59.6%-77.3%), precision of 72.1% (95%CI: 60.1%-82.4%), and an F1 score of 69.9%. The pooled frame-level OOF AUC was 0.726 (95%CI: 0.639-0.806). The pooled OOF ROC curve is shown in Figure 3.

Performance variability and error distribution

Frame-level performance varied across the individual cross-validation folds, with accuracy ranging from 52.1% to 83.0% and balanced accuracy ranging from 51.6% to 83.0%. To characterize the overall frame-level error distribution across the complete cross-validation cohort, a confusion matrix was generated from the

Table 2. Frame-level performance across the 10 folds of the patient-stratified cross-validation

Cross-validation fold	Accuracy	Balanced Accuracy	F1 Score	Precision	Sensitivity	Specificity
1	0.739	0.713	0.640	0.708	0.584	0.842
2	0.736	0.744	0.732	0.667	0.812	0.676
3	0.641	0.628	0.715	0.616	0.854	0.402
4	0.651	0.652	0.590	0.534	0.658	0.647
5	0.830	0.830	0.837	0.798	0.880	0.781
6	0.734	0.732	0.710	0.738	0.684	0.779
7	0.740	0.703	0.616	0.702	0.549	0.857
8	0.616	0.617	0.625	0.606	0.645	0.589
9	0.521	0.516	0.441	0.500	0.395	0.637
10	0.565	0.572	0.574	0.515	0.648	0.496
Pooled OOF	0.682	0.682	0.699	0.721	0.678	0.686

Fold-specific metrics were calculated at the frame level on the respective held-out test sets. Pooled OOF metrics were calculated from all 6,984 held-out frame-level predictions across the ten cross-validation folds and therefore do not represent an arithmetic mean of the fold-specific metrics. OOF: Out-of-fold.

pooled OOF predictions across all ten folds [Figure 4]. Among 3,180 GBM frames, 2,183 (68.6%) were correctly classified and 997 (31.4%) were misclassified as metastasis. Among 3,804 metastasis frames, 2,581 (67.8%) were correctly classified and 1,223 (32.2%) were misclassified as GBM.

Patient-level performance

In addition to the quality-filtered frame-level analysis, patient-level performance was evaluated separately using the annotated frames available for video classification. This video-level analysis was restricted to the same 92 patients included in the final frame-level cohort. Predictions from the annotated frames available for each surgical procedure were aggregated by majority voting. Overall, 66 of 92 patients were correctly classified, corresponding to a patient-level accuracy of 71.7% (95%CI: 61.8%-79.9%; Wilson interval). Among the 46 GBM cases, 33 (71.7%) were correctly classified, and among the 46 brain metastasis cases, 33 (71.7%) were correctly classified. Accordingly, patient-level balanced accuracy, sensitivity, and specificity were each 71.7%. The patient-level ROC-AUC calculated from continuous prediction scores was 0.791 (95%CI: 0.694-0.879).

Metadata-only comparator and sensitivity analyses

The primary metadata-only ridge logistic regression generated OOF predictions for all 92 patients using the same outer 10-fold patient partition. The model correctly classified 65/92 patients (70.7%), with an AUC of 0.779 (95%CI: 0.678-0.869). This was comparable to the continuous patient-level AUC of 0.791 obtained with the multimodal model. Leave-one-out cross-validation yielded similar performance (64/92 patients, 69.6%; AUC 0.764). Location-only sensitivity analyses showed lower performance (63.0% accuracy for lobe alone and 57.6% for lobe plus side), suggesting that the performance of the metadata-only model was not explained solely by tumor location.

Overall patient-level accuracy was similar for the multimodal model (66/92, 71.7%) and the metadata-only comparator (65/92, 70.7%). The individual classifications were discordant in 19/92 patients (20.7%), with 10 patients correctly classified only by the multimodal model and 9 only by the metadata-only model (exact McNemar $P = 1.000$).

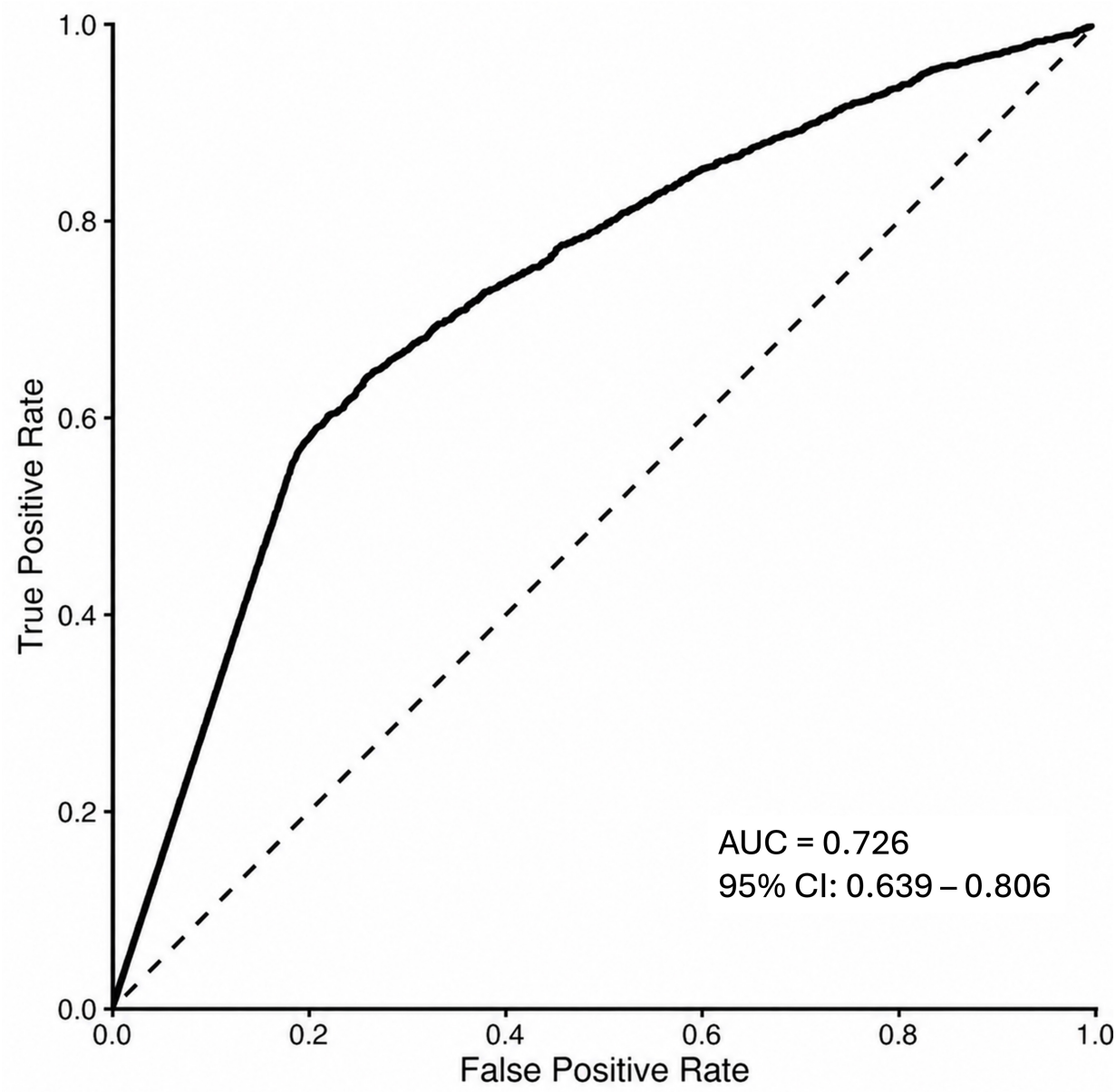


Figure 3. Pooled OOF ROC curve based on 6,984 frame-level predictions from 92 patients across the 10 folds of the patient-stratified cross-validation. The pooled frame-level AUC was 0.726 (95%CI: 0.639-0.806). Each prediction was generated while the corresponding patient was part of the held-out test fold. AUC: Area under the curve; OOF: out-of-fold; ROC: receiver operating characteristic.

Ablation analysis: multimodal vs. image-only model performance

To assess differences in classification performance between the multimodal and image-only approaches, the multimodal model was compared with an image-only model across the same ten cross-validation folds [Table 3]. To ensure a direct comparison, the analysis was restricted to 6,969 frames from 90 patients that were available for both model variants. Across the ten matched folds, mean frame-level accuracy was 67.9% for the multimodal model and 61.2% for the image-only model, corresponding to a mean paired difference of 6.7 percentage points (95%CI: 1.8-11.6; paired *t*-test, $P = 0.013$). Mean balanced accuracy was 67.2% and 61.6%, respectively, with a mean paired difference of 5.6 percentage points (95%CI: 0.6-10.7; $P = 0.032$).

Neurosurgeon diagnostic performance

In order to compare the performance of the AI system to that of human visual assessment, 13 neurosurgeons independently classified the same 33 intraoperative frames from 12 cases as either glioblastoma or brain metastasis. Mean individual reader accuracy across all 13 neurosurgeons was 52.8%, with individual reader

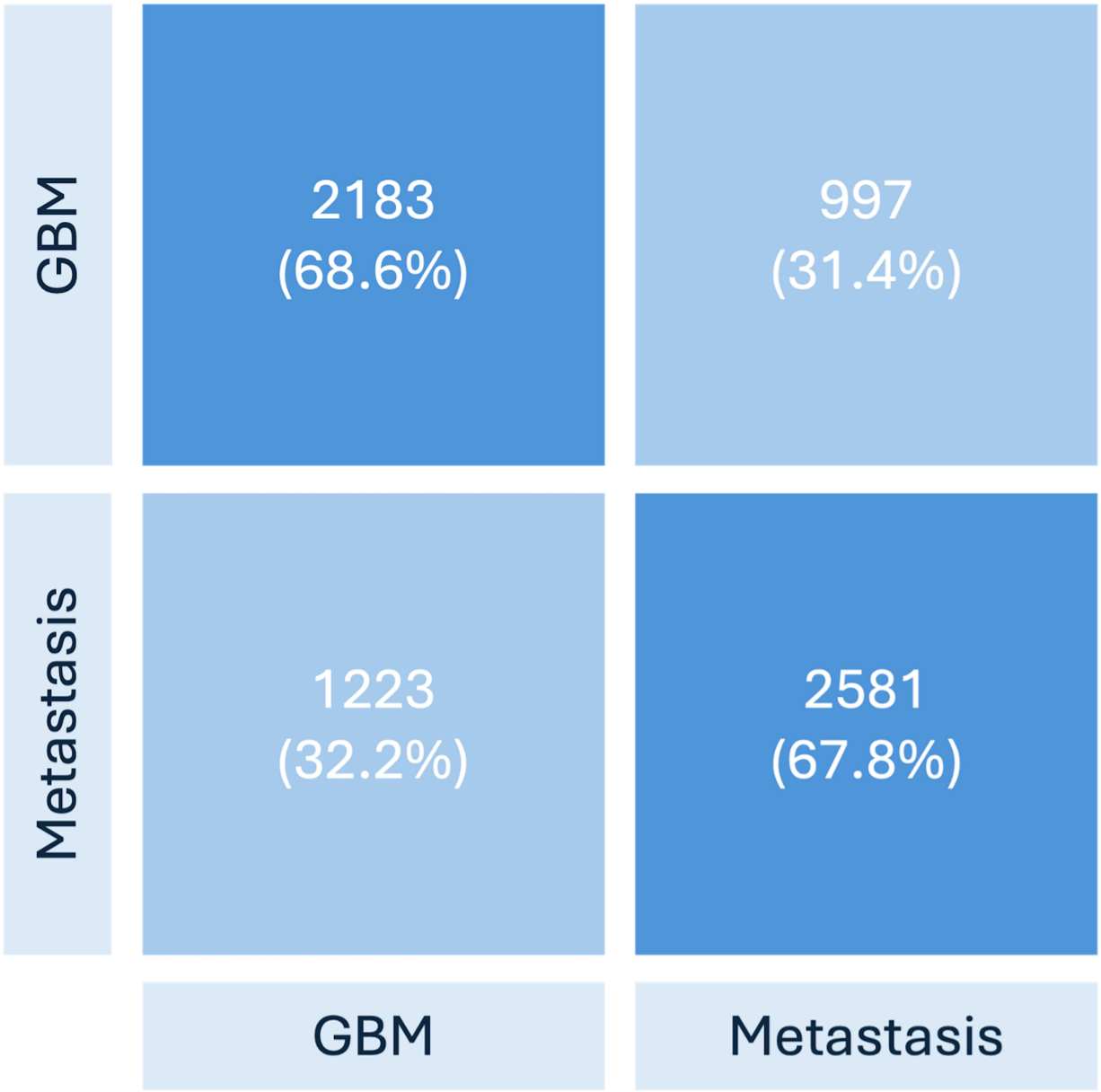


Figure 4. Confusion matrix based on pooled OOF frame-level predictions across all 10 folds of the patient-stratified cross-validation (*n* = 6,984 frames from 92 patients). Cell percentages represent the proportion of frames within each actual tumor class. Overall pooled frame-level OOF accuracy was 68.2%. GBM: Glioblastoma; OOF: out-of-fold.

accuracies ranging from 39.4% to 72.7%. Mean accuracy was 48.5% among the five residents, 51.5% among the four junior board-certified neurosurgeons, and 59.4% among the four experienced board-certified neurosurgeons.

The independently trained hold-out AI model was evaluated at the frame level on the same 33 frames across five independent runs. The model correctly classified 25/33 (75.8%), 24/33 (72.7%), 24/33 (72.7%), 22/33 (66.7%), and 24/33 (72.7%) frames across the five runs, respectively, corresponding to a mean accuracy of 72.1%. This 72.1% accuracy represents the mean performance of the five hold-out AI runs on the 33-frame reader evaluation set and is independent of the image-only model performance reported in the ablation analysis. The AI model therefore showed a descriptively higher mean accuracy than the neurosurgeon readers in this information-restricted visual assessment; no inferential statistical comparison was performed.

Table 3. Fold-wise comparison of multimodal and image-only model performance in the matched ablation analysis

Fold	Multimodal accuracy	Image-only accuracy	Multimodal balanced accuracy	Image-only balanced accuracy
1	0.753	0.620	0.727	0.628
2	0.736	0.615	0.744	0.619
3	0.641	0.724	0.628	0.725
4	0.651	0.593	0.652	0.607
5	0.830	0.767	0.830	0.769
6	0.734	0.598	0.732	0.602
7	0.740	0.723	0.703	0.711
8	0.616	0.525	0.617	0.525
9	0.521	0.502	0.516	0.500
10	0.565	0.454	0.572	0.470
Mean	0.679	0.612	0.672	0.616

The analysis was restricted to 6,969 frames from 90 patients for which predictions from both model variants were available. Accuracy and balanced accuracy are reported for each patient-stratified cross-validation fold.

DISCUSSION

Multimodal intraoperative diagnosis and clinical synergy

Our findings demonstrate the potential of AI-based image classification for intraoperative differentiation of glioblastoma and brain metastases under retrospective experimental conditions. In future clinical applications, AI-based image analysis could potentially complement the surgeon's clinical judgment and established intraoperative diagnostic approaches. Therefore, our approach should be considered a proof-of-concept toward investigating whether visual AI assistance could eventually be incorporated into multimodal intraoperative decision-support systems. Further studies are required to determine how AI-based video analysis could be integrated with established intraoperative diagnostic modalities and whether such integration provides additional clinically relevant information.

Contribution of clinical metadata to multimodal classification

The observed performance difference between the multimodal framework and the vision-only baseline may indicate a potential contribution of clinical metadata to model performance. In the ablation study, mean frame-level accuracy and balanced accuracy were higher for the multimodal model than for the image-only model (67.9% *vs.* 61.2% and 67.2% *vs.* 61.6%, respectively). A metadata-only ridge logistic regression demonstrated discriminative information in the clinical variables, achieving a patient-level accuracy of 70.7%, with a similar result in leave-one-out cross-validation (69.6%). Patient-level ROC-AUCs based on continuous prediction scores were also similar between the multimodal and metadata-only models (0.791 *vs.* 0.779, respectively). Thus, the present data do not demonstrate a clear incremental patient-level accuracy benefit of adding intraoperative image information to metadata alone. Nevertheless, the two approaches did not classify the same patients correctly: 19/92 classifications were discordant, with 10 patients correctly classified only by the multimodal model and 9 only by the metadata-only model. These findings indicate different individual classification patterns but do not establish superiority or complementary clinical benefit of either approach. Location-only sensitivity analyses yielded lower performance than the full metadata model (63.0% accuracy for lobe alone and 57.6% for lobe plus side), suggesting that the metadata-only performance was not explained solely by tumor location.

Synergy between human expertise and AI-based diagnostics

Our initial hypothesis was that experienced surgeons develop an intuitive sense for tumor entities during resection. However, the reader assessment demonstrated that differentiation based on static intraoperative images alone was challenging even for experienced neurosurgeons. Mean individual reader accuracy across

all 13 neurosurgeons was 52.8%, ranging from 39.4% to 72.7%. Mean accuracy increased descriptively with clinical experience, from 48.5% among residents to 51.5% among junior board-certified neurosurgeons and 59.4% among experienced board-certified neurosurgeons. These results may in part reflect the information-restricted design of the reader assessment, in which surgeons were provided with static images without dynamic video input, clinical or radiological information, or other sensory information such as haptic feedback regarding tissue consistency. On the same 33-frame evaluation set, the independently trained hold-out AI model achieved a mean accuracy of 72.1% across five runs, which was descriptively higher than the mean reader accuracy of 52.8%. No inferential statistical comparison between AI and human performance was performed. The observed difference suggests that the model may extract visual information from static intraoperative images that is not readily captured by human visual assessment under these information-restricted conditions. However, the reader comparison was conducted on a small, manually selected best-case image set and should not be interpreted as evidence of superiority over surgeon assessment in routine clinical practice. Beyond tumor classification, similar applications of AI-based image analysis have been investigated in other surgical specialties, including colorectal surgery, where computer vision approaches have shown potential for recognizing anatomical structures and assisting surgical navigation[14,18]. In diagnostic imaging, AI-based image analysis has shown potential to improve objectivity and reproducibility in selected radiological diagnostic tasks[14].

Potential benefits and clinical standardization

Future integration of such an AI model into a surgical microscope could potentially provide additional objective information during surgery. However, the present study did not evaluate inference latency, continuous video-stream processing, actual microscope integration, workflow effects, or clinical decision-making. These aspects require prospective technical and clinical evaluation before conclusions regarding workflow efficiency or intraoperative utility can be drawn. More generally, AI-based image analysis has the potential to provide standardized assessments based on predefined computational criteria. Human image interpretation can show inter-reader variability and may be influenced by factors including reader experience and fatigue[19,20]. In contrast, a trained AI model consistently applies the same computational criteria to each input image under otherwise unchanged conditions. Whether this reproducibility translates into more standardized intraoperative assessment across surgeons or institutions remains to be established through prospective multicenter validation.

Architectural considerations and constraints for clinical translation

The architecture was designed with computational efficiency in mind to facilitate potential future intraoperative implementation. In a clinical setting, local execution of AI models may be advantageous because it reduces dependence on external network connectivity and may facilitate integration into medical devices[21]. Computational efficiency was therefore considered during model development as a potential requirement for future local implementation. Accordingly, the present model was designed to balance classification performance with computational complexity. However, inference latency and continuous-stream performance were not formally measured in the present study, and the model was not directly integrated into a surgical microscope. Real-time capability and technical feasibility within an actual intraoperative workflow therefore remain to be established in dedicated prospective implementation studies. Importantly, the present proof-of-concept study did not evaluate a predefined clinical decision pathway or the consequences of model-guided surgical decisions. Therefore, no specific intraoperative decision or modification of surgical strategy can currently be assigned to the model output. In the present retrospective study, model predictions were generated offline without intraoperative user interaction. For potential future clinical use, the extent of required user interaction, including image selection and interpretation of model outputs, as well as the necessary level of user expertise, has not yet been established. Procedures for handling inadequate input images also remain to be defined in prospective implementation studies. Further model

development using larger and more diverse datasets, independent external validation, and technical implementation on suitable intraoperative hardware are required before a clinical intended use can be defined and prospectively evaluated.

Limitations

The relatively small study cohort of 92 patients, retrospective single-center proof-of-concept design, and reliance on internal cross-validation without independent external validation limit the generalizability and clinical applicability of our findings. Given the cohort size relative to model complexity, overfitting remains an important concern and may have contributed to the variability in performance across cross-validation folds. Furthermore, hyperparameter selection without a separate validation set may have introduced optimistic bias into the reported performance estimates. Preoperative radiological diagnoses were unavailable for 15 patients, and diagnostic concordance was assessed retrospectively using broad diagnostic categories, precluding direct comparison with AI classification accuracy. No formal subgroup-specific performance or algorithmic fairness analyses were conducted because further stratification of the 92 patients would have resulted in small subgroups and potentially unstable estimates. Consequently, potential performance differences across demographic and clinical subgroups, including age and sex, remain unknown. Larger, more diverse external cohorts are needed to assess generalizability, subgroup performance, and potential algorithmic biases.

Exclusion of videos with insufficient image quality, including excessive bleeding, blurring, or poor tumor visibility, may have introduced selection bias and contributed to performance variability. Moreover, manual selection of informative frames created an enriched experimental setting that does not reflect automated analysis of unselected continuous operative video. Image quality may further affect classification performance through bleeding, blurring, overexposure, poor illumination, and limited visibility of reference tissue. Sampling errors and tumor heterogeneity represent additional potential sources of bias. The favorable images selected for the comparative reader assessment constituted a best-case scenario that may not reflect routine intraoperative conditions. Inference latency, continuous-stream performance, and direct microscope integration were not evaluated. Accordingly, real-time feasibility, workflow efficiency, and effects on surgical decision-making, surgical precision, or patient outcomes remain unestablished. The clinical consequences of incorrect classifications were also not assessed and will depend on how model outputs are incorporated into surgical decision-making. The model's specific intended clinical use therefore remains to be defined, and prospective technical and clinical evaluation is required before implementation.

The metadata-only comparator was internally validated in the same small single-center cohort and may have been affected by sparse categorical levels, particularly the imbalance in cerebellar tumor location. Its generalizability and any incremental benefit of the multimodal model therefore require external validation. Model calibration was not formally assessed because the study focused on classification performance rather than the reliability of predicted probabilities. Calibration should be evaluated before probability-based clinical decision support is considered. Finally, the paired *t*-tests used in the ablation analysis should be interpreted as exploratory because cross-validation folds share training data and are not statistically independent.

Conclusion

This proof-of-concept study demonstrates the technical feasibility of using multimodal AI-based analysis of intraoperative microscopic images and clinical metadata to differentiate glioblastoma from brain metastases. The observed frame- and patient-level performance supports further investigation of this approach in the context of intraoperative tumor assessment. Importantly, the addition of intraoperative microscope images did not demonstrate an incremental improvement in patient-level classification accuracy compared with

clinical metadata alone. Future studies should expand the dataset through large-scale prospective acquisition of intraoperative video data, include a broader range of tumor entities and clinical settings, and perform independent external validation. In addition, prospective studies are required to determine whether the model can be successfully integrated into the intraoperative workflow and whether its predictions provide clinically meaningful decision support. The effects of such integration on surgical strategy, workflow efficiency, surgical outcomes, and patient outcomes remain to be established.

DECLARATIONS

Authors' contributions

Conceptualization, supervision, annotation, writing - review & editing: Dorfer C
Data curation, investigation, annotation, methodology, formal analysis, visualization, writing - original draft: Makolli J
Methodology, software, formal analysis, writing - review & editing: Bacher N, Saur S
Methodology, software, writing - review & editing: Spier N
Methodology, radiological assessment, writing - review & editing: Mallouhi A
Had access to the anonymized study data and performed and verified the final analyses: Makolli J, Bacher N, Saur S, Spier N
Retained final authority over the interpretation of the study findings and the decision to submit the manuscript for publication: Dorfer C

Availability of data and materials

Datasets generated during the current study are available from the corresponding author upon reasonable request. Determination of a reasonable request and subsequent data provision are subject to institutional approval. The study protocol is available from the corresponding author upon reasonable request. The academic authors had access to the clinical study data and all model predictions and analysis outputs underlying the results reported in this manuscript.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool OpenAI's ChatGPT (version GPT-5, released 2025-08-07) was used solely to improve language clarity and grammar. The video abstract accompanying this article was created using Higgsfield, an AI-based video generation tool. These tools did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This study was supported in part by Carl Zeiss Meditec AG through a financial contribution associated with contractual rights to use the research data and results within a joint research collaboration with the Medical University of Vienna. The academic research team was responsible for the study design, clinical data collection, scientific analysis, interpretation of the findings, and manuscript preparation. Carl Zeiss Meditec AG/Carl Zeiss AG provided technical expertise and support where required, including the development of the AI model, provision of the annotation tool, and generation of model predictions and analysis outputs.

Conflicts of interest

Bacher N, Saur S, and Spier N are employees of Carl Zeiss Meditec AG/Carl Zeiss AG. Carl Zeiss Meditec AG/Carl Zeiss AG developed the AI model evaluated in this study and provided the annotation tool used for image annotation, constituting in-kind contributions to the study. The academic research team provided the clinical metadata and intraoperative video recordings to ZEISS for model development. The three ZEISS-affiliated co-authors contributed to model development and the generation of model predictions and analysis outputs. The other authors declare no conflicts of interest.

Ethical approval and consent to participate

This study was approved by the Ethics Committee of the Medical University of Vienna (Protocol No. 1250/2024). The study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all patients for the recording of intraoperative videos and their use for scientific research purposes.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Knutsson M, Salomonsson T, Durmo F, et al. Differentiation between glioblastoma and solitary brain metastases using perfusion and amide proton transfer weighted MRI. *Front Neurosci.* 2025;19:1533799. DOI PubMed PMC
2. Cao X, Tan D, Liu Z, et al. Differentiating solitary brain metastases from glioblastoma by radiomics features derived from MRI and 18F-FDG-PET and the combined application of multiple models. *Sci Rep.* 2022;12:5722. DOI PubMed PMC
3. Que T, Huang G, Tan J, et al. Supramaximal resection based on en-bloc technique reduces tumor burden and prolongs survival in primary supratentorial lobar glioblastoma. *J Neurooncol.* 2023;164:557-68. DOI PubMed
4. Al-Holou WN, Hodges TR, Everson RG, et al. Perilesional resection of glioblastoma is independently associated with improved outcomes. *Neurosurgery.* 2020;86:112-21. DOI PubMed PMC
5. Altieri R, Corvino S, La Rocca G, et al. Clinical implication of brain metastases en-bloc resection: surgical technique description and literature review. *J Pers Med.* 2024;14:1110. DOI PubMed PMC
6. Byun YH, Won JK, Hong DH, et al. A prospective multicenter assessor blinded pilot study using confocal laser endomicroscopy for intraoperative brain tumor diagnosis. *Sci Rep.* 2024;14:6784. DOI PubMed PMC
7. Romeo E, Markopoulos G, Voulgaris S, Vartholomatos G, Alexiou GA. The role of intraoperative flow cytometry on intracranial tumor surgery : a scoping review. *Neurosurg Rev.* 2025;48:637. DOI PubMed
8. Malinova V, von Eckardstein K, Mielke D, Rohde V. Diagnostic yield of fluorescence-assisted frame-based stereotactic biopsies of intracerebral lesions in comparison with frozen-section analysis. *J Neurooncol.* 2020;149:315-23. DOI PubMed
9. Harms JWA, Streckert EMS, Kiolbassa NM, et al. Confounders of intraoperative frozen section pathology during glioma surgery. *Neurosurg Rev.* 2023;46:286. DOI PubMed
10. Plesec TP, Prayson RA. Frozen section discrepancy in the evaluation of central nervous system tumors. *Arch Pathol Lab Med.* 2007;131:1532-40. DOI PubMed
11. Movahed-Ezazi M, Nasir-Moin M, Fang C, et al. Clinical validation of stimulated raman histology for rapid intraoperative diagnosis of central nervous system tumors. *Mod Pathol.* 2023;36:100219. DOI PubMed PMC
12. Fava A, di Russo P, Morace R, Esposito V. Anatomical landmarks during high-magnification microsurgery for a safe and effective resection of high-grade gliomas: how I do it. *Acta Neurochir.* 2023;165:4235-40. DOI PubMed
13. Sauvigny T, Ricklefs FL, Hoffmann L, Schwarz R, Westphal M, Schmidt NO. Features of tumor texture influence surgery and outcome in intracranial meningioma. *Neurooncol Adv.* 2020;2:vdaa113. DOI PubMed PMC
14. Xie Y, Wang Z, Zeng Z, et al. Applications of artificial intelligence in cancer diagnosis and treatment. *Cancer Med.* 2026;15:e72143. DOI PubMed PMC
15. Faraj CA, Young CC, Beckham TH, Weinberg JS. Intraoperative technological advances and new frontiers in precision glioma surgery. *J Neurooncol.* 2026;178:89. DOI PubMed PMC
16. Kiyasseh D, Ma R, Haque TF, et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat Biomed Eng.* 2023;7:780-96. DOI PubMed PMC
17. Azad R, Kazerouni A, Heidari M, et al. Advances in medical image analysis with vision Transformers: a comprehensive review. *Med Image Anal.* 2024;91:103000. DOI PubMed
18. Quero G, Mascagni P, Kolbinger FR, et al. Artificial intelligence in colorectal cancer surgery: present and future perspectives. *Cancers.* 2022;14:3803. DOI PubMed PMC
19. Zhan H, Schartz K, Zygmont ME, Johnson JO, Krupinski EA. The impact of fatigue on complex CT case interpretation by radiology residents. *Acad Radiol.* 2021;28:424-32. DOI PubMed

20. Hsieh SS, Cook DA, Inoue A, et al. Understanding reader variability: a 25-radiologist study on liver metastasis detection at CT. *Radiology.* 2023;306:e220266. DOI PubMed PMC
21. Walji A, Masino C, Williams S, et al. Optimizing AI implementation for surgery: recommendations for infrastructure and deployment in the operating room. *NPJ Digit. Med.* 2026;9:621. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.