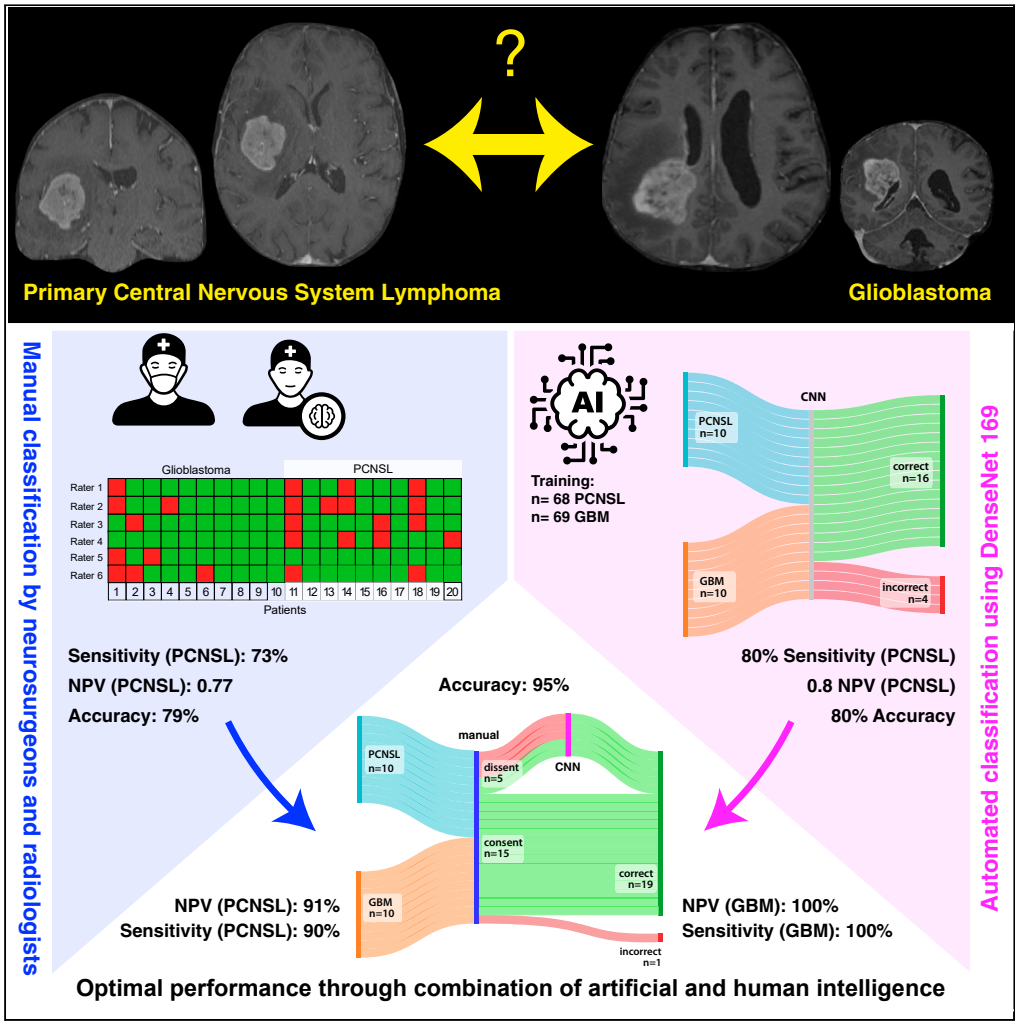
iScience

CellPress
OPEN ACCESS

Article

Deep learning aided preoperative diagnosis of primary central nervous system lymphoma



Paul Vincent Naser, Miriam Cindy Maurer, Maximilian Fischer, ..., Sandro M. Krieg, Peter Neher, Jan-Oliver Neumann

paul.naser@med.uni-heidelberg.de

Highlights

Identifying PCNSL pre-surgery is challenging, even for skilled radiologists

We developed and tested an automated workflow to diagnose PCNSL based on MRI

Automated and manual classification were comparably accurate at approx. 80%

Combining human and artificial intelligence yielded the optimal diagnostic performance

Naser et al., iScience 27, 109023
February 16, 2024
https://doi.org/10.1016/j.isci.2024.109023

iScience

CellPress
OPEN ACCESS

Article

Deep learning aided preoperative diagnosis of primary central nervous system lymphoma

Paul Vincent Naser,[1,2,6,10,12,13,*] Miriam Cindy Maurer,[3,11,12] Maximilian Fischer,[2,3,7] Kianush Karimian-Jazi,[2,4] Chiraz Ben-Salah,[2,5] Awais Akbar Bajwa,[1,2] Martin Jakobs,[1,2,6] Christine Jungk,[1,2] Jessica Jesser,[2,4] Martin Bendszus,[2,4] Klaus Maier-Hein,[3,7,8,9,10] Sandro M. Krieg,[1,2] Peter Neher,[3,7,9] and Jan-Oliver Neumann[1,2,6]

SUMMARY

The preoperative distinction between glioblastoma (GBM) and primary central nervous system lymphoma (PCNSL) can be difficult, even for experts, but is highly relevant. We aimed to develop an easy-to-use algorithm, based on a convolutional neural network (CNN) to preoperatively discern PCNSL from GBM and systematically compare its performance to experienced neurosurgeons and radiologists. To this end, a CNN-based on DenseNet169 was trained with the magnetic resonance (MR)-imaging data of 68 PCNSL and 69 GBM patients and its performance compared to six trained experts on an external test set of 10 PCNSL and 10 GBM. Our neural network predicted PCNSL with an accuracy of 80% and a negative predictive value (NPV) of 0.8, exceeding the accuracy achieved by clinicians (73%, NPV 0.77). Combining expert rating with automated diagnosis in those cases where experts dissented yielded an accuracy of 95%. Our approach has the potential to significantly augment the preoperative radiological diagnosis of PCNSL.

INTRODUCTION

Patients with rapidly progressing intracranial tumors are often initially admitted because of neurological deficits, such as paresis, stemming from the mass effect caused by the lesion and interfering with normal brain function. Antiedematous therapy using corticosteroids such as dexamethasone provides effective acute symptom relief to those patients suffering from elevated intracranial pressure. One of the most common malignant intraaxial tumors is glioblastoma (GBM), accounting for nearly half of all newly diagnosed intracranial malignancies.[1] Primary central nervous system lymphomas (PCNSL), in contrast, are exceedingly rare lesions, constituting between 4 and 7% of newly diagnosed intracranial lesions.[2,3] MR-morphologically, GBM and PCNSL can share a range of morphological features, such as large perifocal edema, inhomogeneous contrast enhancement, and, in some cases, central necrosis, hindering an accurate preoperative diagnosis.[4,5] In these cases especially, stereotactic biopsy is the gold standard for histopathological analysis and is associated with very low perioperative morbidity and mortality.[6]

Clinically relevant, in cases of PCNSL, treatment with corticosteroids severely hinders pathological diagnosis,[7] which remains absolutely necessary.[8,9] Additionally, stereotactic biopsy is a specialized procedure only available in large centers, putting neurologists and neurosurgeons in peripheral hospitals or emergencies in the conflict, whether to treat patients with corticosteroids for symptom relief, risking inconclusive pathology results in case of PCNSL or forcing patients to endure the debilitating symptoms of cerebral edema until a biopsy can be performed.

[1]Heidelberg University Hospital, Department of Neurosurgery, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
[2]Heidelberg University, Medical Faculty, Grabengasse 1, 69117 Heidelberg, Germany
[3]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[4]Heidelberg University Hospital, Department of Neuroradiology, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
[5]Department of Diagnostic and Interventional Radiology, University Hospital Heidelberg, Im Neuenheimer Feld 110, 69120 Heidelberg, Germany
[6]Heidelberg University Hospital, Division of Stereotactic Neurosurgery, Department of Neurosurgery, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
[7]German Cancer Consortium (DKTK), partner site Heidelberg, Heidelberg, Germany
[8]National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and the University Medical Center Heidelberg, 69120 Heidelberg, Germany
[9]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, 69120 Heidelberg, Germany
[10]AI Health Innovation Cluster, German Cancer Research Center (DKFZ), Heidelberg, Germany
[11]Department of Medical Informatics, University Medical Center Göttingen, Von-Siebold-Straße 3, 37075 Göttingen, Germany
[12]These authors contributed equally
[13]Lead contact
*Correspondence: paul.naser@med.uni-heidelberg.de
https://doi.org/10.1016/j.isci.2024.109023

Advances in biomedical image processing have impressively proven the versatile use cases for computer programs to analyze and evaluate medical images.[10] The current study, therefore, aimed to develop an automated diagnostic workflow to distinguish between PCNSL and GBM, to help clinicians decide which patients to administer antiedematous corticosteroid therapy.

RESULTS

Tumor segmentation

Automated tumor segmentation with manual vetting yielded acceptable results, with comparable sizes of contrast-enhancing and non-enhancing tumor volumes, as well as edema volumes across the training and test groups. No statistical differences were detected regarding the tumor volumes, both when comparing GBM to PCNSL, or the respective training and test groups (Figure 1).

Manual tumor classification

Overall accuracy of tumor prediction by expert assessment was high at 79.16% (95/120 predictions). In the PCNSL group, an insignificantly higher prediction accuracy (49/60) was observed (GBM 46/60; $p = 0.5$, Chi-square). Details about the accuracies and negative/positive predictive values are given in Table S1. Analysis of the receiver operating characteristics curve (ROC) (Figure S2D), considering the reported confidence in the diagnosis, revealed an area under the curve (AUC) of 0.8091 for all predictions, 0.9 for PNCSL, and 0.603 for GBM, respectively.

Linear regression analysis was used to analyze the effect of the rater's experience in diagnosing cMRI on their performance. Experience neither significantly correlated with the self-assessed PCNSL prediction accuracy ($F = 0.44$, $p = 0.56$) nor prediction absolute accuracy for GBM ($F = 0.59$, $p = 0.48$) or PCNSL ($F = 0.08$, $p = 0.79$) (Figure 2). Time to prediction took 110.3 ± 53 s but was significantly variant across raters (Figure S2E). Fitting a mixed-effects model with Šídák's post-hoc test comparing the time to diagnosis of radiologists and neurosurgeons, a highly significant effect was discovered in patient's heterogeneity ($p = 0.0005$), but not in the training of the rater ($p = 0.74$). No significant difference in time expenditure was detected across all raters between correct and incorrect diagnoses (106.4 ± 52s vs. 124.8 ± 55.49; $p = 0.5$). Likewise, no significant difference was detected between PCNSL and GBM cases (114.9 ± 51.8 s vs. 105.8 ± 54.3 s, $p = 0.93$).

As part of the assessment questionnaire, experts rated tumor aspects (contrast-enhancing, non-enhancing, necrosis, and edema), the tumor volume, and the available imaging modalities numerically based on the relevance to their diagnosis. T1-CE was overwhelmingly deemed the most important MR modality (Figure 2). The second most important modality was fluid attenuated inversion recovery (FLAIR), followed by T2. Native T1 was assessed as least important. No significant difference between the GBM and PCNSL groups was noted for any of these modalities assessed by Chi-square testing.

When asked to rate the relevance of different tumor aspects, human raters attributed the most importance to the contrast-enhancing parts of the tumor in both PCNSL and GBM. Secondary, in comparison to PCNSL, raters attributed significantly more importance to necrosis ($p = 0.017$, Chi-square) in GBM patients, whereas the relative importance of the non-contrast-enhancing aspects as well as of the edema did not differ (both $p > 0.05$, Chi-square).

Raters scored the importance of tumor volumes on a percentual scale from 0 to 100%. Significant differences were detected between raters, however, when examining all raters combined and individually, the tumor size did not significantly correlate with an individual's assessment of its importance (Figure 2).

After giving the raters access to all imaging sequences, the accuracy increased to 86.67%. However, some raters, in fact, miscorrected their initially accurate diagnosis (Figure S2A). Raters were asked, which sequence beyond the four provided in the first stage was most important for the diagnosis. For PCNSL, raters found diffusion sequences (apparent diffusion coefficient, ADC and diffusion-weighted imaging, DWI) most insightful ($n = 22$), followed by susceptibility-weighted images (SWI) ($n = 6$) and cerebral blood volume sequences (CBV, $n = 2$). In GBM, a trend toward more frequently requesting CBV ($n = 7$) was found, whereas both SWI ($n = 5$) and diffusion-weighed images were likewise deemed helpful ($n = 14$). The difference between the groups was insignificant ($p = 0.117$, X^2).

Development of the automated classification algorithm

Two general approaches (ResNet and DenseNet) were initially investigated. Both models were trained using 5-fold internal cross-validation and ensembled before testing. The first experiment was conducted to test the depth of ResNet. ResNet10, ResNet18, ResNet34, ResNet50, and ResNet101 were evaluated, and hyperparameter tuning was initiated for the best-performing ResNet10 and ResNet18. DenseNet169 and DenseNet264 were likewise subjected to hyperparameter optimization. Comparing these four neural networks, DenseNet169 performed best in the grid search cross-validation and was used in the automated predictions of tumor entities (Table 1).

Automated classification

After training the neural network, the data from hitherto unknown patients (test group) were presented to the algorithm, which provided a binary diagnosis of either GBM or PCNSL at a decision threshold of 0.5. The optimized DenseNet169 correctly diagnosed the entity in 80% of cases (16/20), with two PCNSL and two GBM cases each misdiagnosed (Figure 2F). The ROC-AUC was 0.9, and the positive and negative predictive values were 0.8, respectively. The mean time to prediction was 0.045 ± 0.0003 s and was not significantly different between correct/incorrect predictions and GBM/PCNSL (Figures S2B and S2C). We assessed the importance that the neural network ascribed to certain regions of the images using saliency maps (examples in Figures 3G–3I). When comparing true and false predictions, no significant differences were detected overlap between the saliency maps and the segmented tumor areas ($p = 0.345$, unpaired t-test).

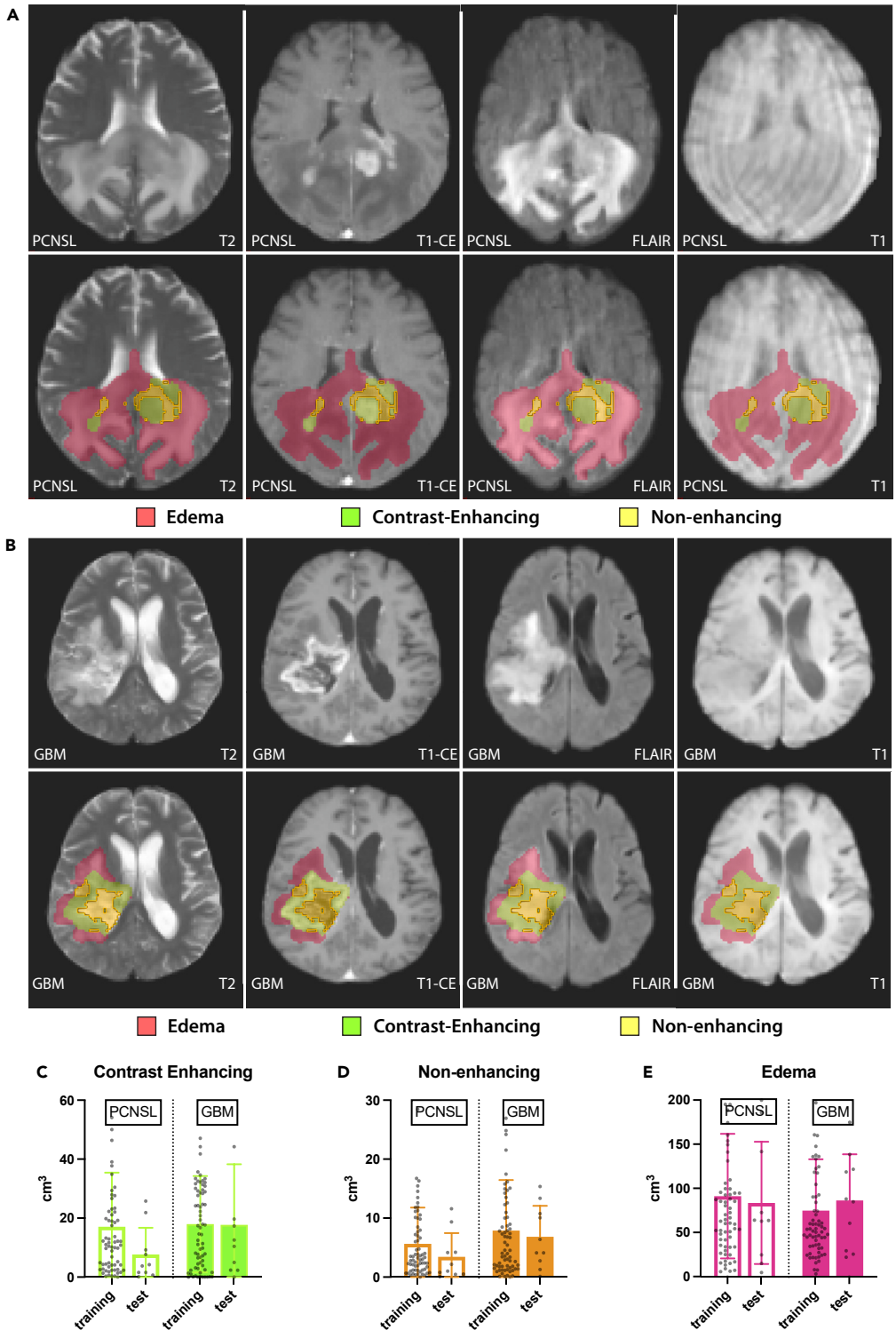


Figure 1. Results of automated brain tumor segmentation

Examples of automated tumor segmentation results are (A) (PCNSL) and (B) (GBM). From left to right, T2, T1-CE, FLAIR, and T1 sequences without (top row) and with (bottom row) mask.

(C–E) illustrate the absolute volumetric values for contrast-enhancing, non-enhancing tumor parts and the perifocal edema. Bars represent mean ± S.D. n = 79 for GBM (69 training/10 testing) and 78 for PCNSL (68 training/10 testing). ANOVA testing confirmed no significant differences between the groups.

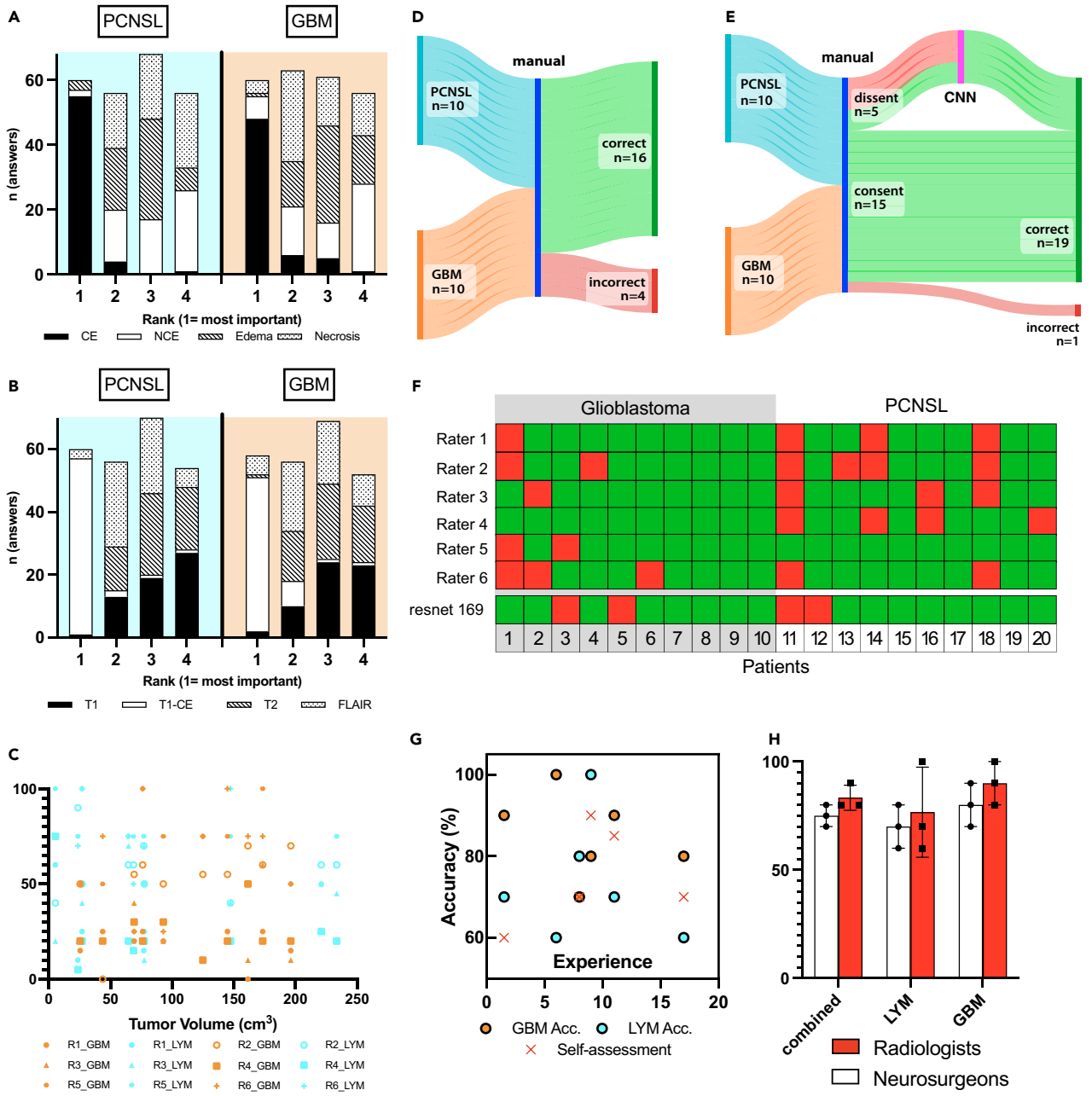


Figure 2. Expert diagnosis of PCNSL and GBM

All raters ascribed high levels of importance to the contrast-enhancing aspects of the tumor (A) and the T1CE MR-sequence (B). Neither when comparing all raters nor each rater individually a significant correlation between tumor size and the attributed importance of size for the diagnosis was detected (C). Sankey chart illustrating the accuracy of human raters (D), and the increased accuracy (95%) when utilizing the CNN for the cases where human raters dissented (E).The matrix in (F) depicts correct (green) and incorrect (red) diagnoses for all raters and the CNN. No significant correlation was found between the absolute experience raters reported with cMRI and accuracy in diagnosing GBM/PCNSL or the self-assessed accuracy (G). No significant difference was detected in prediction accuracy when comparing neurosurgeons with radiologists (H). Bars in (H) represent mean ±S.D.

Importance of MRI sequences

As a similar rating of the importance of the MRI modalities was technically impossible, several experiments were conducted to evaluate the aspects most important to the network. Firstly, the algorithm was only presented with one modality (Figure 3D). The network only provided with T1CE performed worst (AUC 0.47), whereas only presenting T1 yielded in comparison the best results (AUC 0.65), although still significantly less accurate than training with all modalities (AUC0.9). In a second experiment, iteratively, modalities were removed from the network.

Table 1. Overview of evaluated neural networks subjected to hyperparameter tuning

Model	Dropout	AUG	Batch size	Best Score
ResNet10	no	No	20	0.6242
ResNet18	no	Yes	12	0.6242
DenseNet169	yes	Yes	14	0.6368
DenseNet264	no	No	14	0.6220

All models were evaluated with a learning rate of 0.0001 over 150 epochs. DenseNet169 showed the best performance and was further used in automated tumor prediction.

Interestingly, removal of the FLAIR dataset showed the highest adverse effect, with the AUC dropping from 0.9 to 0.71, whereas removing any other single imaging modality had minor effects. In a final experiment, to elucidate the effect of including the segmentation masks on the accuracy of the predictions, this feature was removed. Of note, removing the predefined masks merely lowered the AUC from 0.9 to 0.84, suggesting this preprocessing did not contribute notably to the network's performance.

Integration of manual and automated classification

The automated and manual classification was equally correct, with accuracies of 80% and 79%, respectively. Interestingly, the tumors incorrectly identified differed between the two (Figure 2F). The convolutional neural network (CNN) classified those tumors correctly, which human raters disagreed about (>1 discerning diagnosis from the majority, Figure 2D). We performed a layered analysis, in which only those tumors experts dissented about were tested by the neural network. By combining manual and automated diagnosis, an accuracy of 95% was achieved (Figure 2E).

DISCUSSION

MR-morphologically, PCNSL and high-grade gliomas like GBM share many common characteristics, such as large perifocal edemas, contrast enhancement, and irregular structural composition. Indeed, preoperative diagnosis can be challenging even for experienced radiologists. One recent study evaluated the performance of two experienced radiologists in 93 PCNSL and 48 GBM cases, reporting accuracies of 74.2% and 82.9%, respectively. These findings closely resemble our data, with our cumulative accuracy of nearly 80% (75–90%). The literature often cites extensive experience in diagnosing PCNSL as crucial to the diagnosis; however, to our best knowledge, no study systematically evaluated the performance of junior vs. senior physicians in this aspect. Our data did not support the notion of long-term experience being a decisive factor in the accurate preoperative diagnosis of PCNSL, as experience did not correlate with accuracy. Of note, we also compared the performance of radiologists with that of neurosurgeons and found no significant differences in the diagnostic accuracies or time needed to arrive at the final diagnosis. When expert raters had access to all imaging modalities, the accuracy increased slightly, however, some clinicians disimproved their initially correct diagnosis.

This further supports the notion, that while many cases can be easily diagnosed, in special cases, even experts disagree (Figure 2F). To address these special cases, several decision trees and guidelines have been developed.[11] Further, special MRI sequences or the incorporation of PET-CT in the diagnostic workflow increased accuracy.[12,13]

However, none of these techniques have gained widespread acceptance. In these difficult cases, an unbiased, objective measure judging the likelihood of a given tumor being PCNSL promises to help clinicians. A recent systematic review lists 23 previous studies addressing this question using a form of machine learning.[14] The vast majority of studies relied on training traditional machine learning algorithms with radiomics data. Among these, extracting radiomics features and fitting a logistic regression model reportedly achieved accuracies up to 91.2%.[14–16] A common point of criticism in these studies is the relative underrepresentation of PCNSL when compared to GBM, which, inherently, is due to the scarcity of cases.[14] However, training neuronal networks on imbalanced datasets can severely skew the model's performance.[17]

Recent years have impressively shown the versatile applications CNNs can have in biomedical image analysis.[10,18] Of the 23 studies reviewed by Petersen and colleagues, only two studies employed this technology to the question of PCNSL/GBM distinction.[14] Yamashita and colleagues trained a neural network with 126 cases, among those 58 high-grade gliomas and 12 PCNSL. Their methodology yielded an impressive 0.949 AUC. However, it should be noted that, owing to the small sample size, no external validation, but the internal leave-one-out method was utilized, potentially making this approach susceptible to over-fitting.[19]

Using single images extracted from 320 T1-CE datasets, McAvoy and colleagues trained several neural networks based on the EfficientNetB4 architecture.[18] Their approach consisted of manually extracting axial slices containing tumors and feeding the downscaled 2D images into the classifier.[18] While the reported AUC values were high (0.93–0.95), the authors admit a high amount of data loss in the preselection and preprocessing steps, potentially biasing their results.[18]

Park and colleagues developed an autoencoder inferring perfusion from experimental dynamic susceptibility contrast MRI datasets.[20] They could discern the two entities with an AUC of 0.93 in their external validation set.[20]

Recently, Tariciotti and colleagues developed a ResNet101 model on 121 patient datasets, including 47 GBM, 37 metastasis, and 37 PCNSL.[21] The data were split 70:30 in training and test datasets. Their workflow included manual tumor segmentation and supplying only

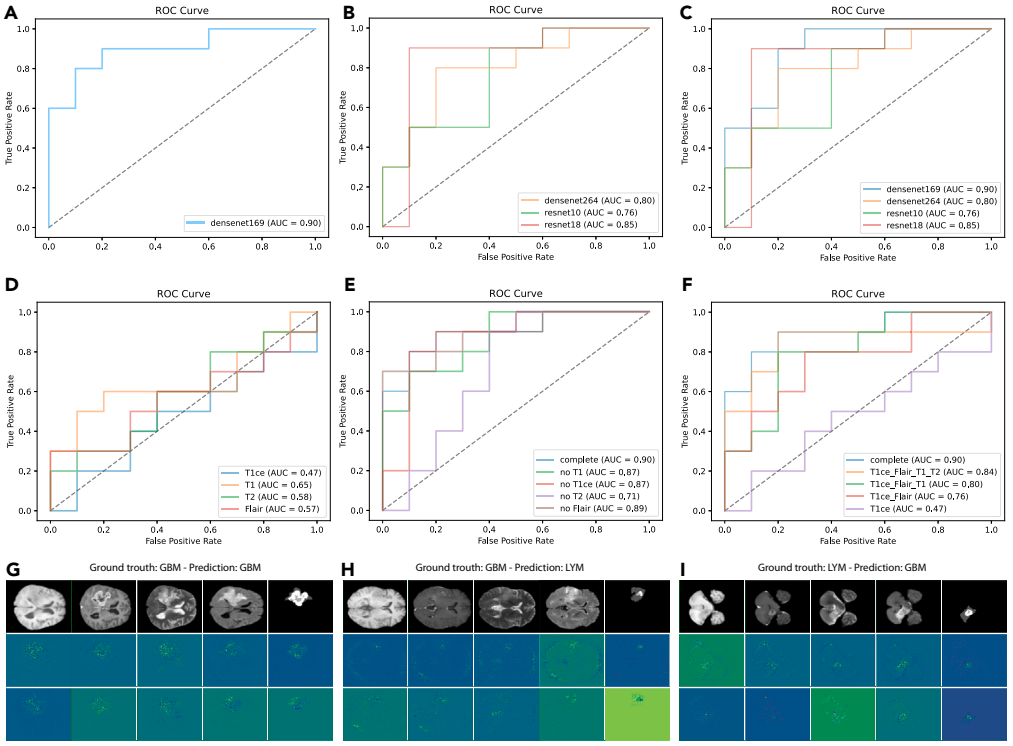


Figure 3. Automated diagnosis of GBM and PCNSL

(A) The ROC-AUC curve of the best-performing neural network (densenet169) with an AUC of 0.9 on the test dataset (n = 20). Other approaches were dismissed for their inferior performance (B + C).

(D) When only evaluating one modality, neural network performance was severely diminished.

(E) When removing one modality from the network, removing T2 exerted the strongest effect.

(F) Removing the segmented tumor masks only lightly affected the performance of DenseNet.

(G–I) Illustrations of saliency maps visualizing the importance assigned to image areas. The top row depicts from left T1, T1CE, T2, FLAIR, and the segmented tumor masks. The middle and bottom rows show saliency maps generated through the integrated gradient (middle) or guided backpropagation (bottom).

the segmented regions of interest in 2D matrices to the CNN.[21] PCNSL detection accuracy was notably high at 94.65%, whereas GBM was detected with an accuracy of 83.08%, however, the groups differed in tumor volume with significantly smaller PCNSL included, and the number of evaluated patients was low.[21] While these discussed studies proved the possibility of predicting PCNSL with high accuracy, using their methodology heavily relies on either special image modalities acquired,[20] or laborious manual tumor segmentation, prohibiting clinical use of these technologies.[21]

We, therefore, set out to develop a deep-learning algorithm based on a CNN for the preoperative distinction of GBM from PCNSL, which (1) does not require special MR-sequences beyond the standard T1, T1CE, T2, and FLAIR sequences and (2) is out-of-the-box useable on MR data without laborious preprocessing, manual segmentation, or image preselection.

The model developed here could accurately predict the presence of GBM or PCNSL in 80% of cases, with an NPV for PCNSL of 0.8, slightly higher than our expert group. Expectedly, the automated diagnosis was vastly faster by a factor of 2000. Interestingly, the network correctly predicted those cases where radiologists and neurosurgeons had the lowest accuracy but failed to predict cases unequivocally diagnosed by human raters. It is unclear whether this discrepancy is purely coincidental or whether the neuronal network considered completely different parameters than expert physicians. Analysis of the net's saliency did not yield significant differences between correct and incorrect predictions, suggesting that, indeed, the algorithm took into account the relevant tumor areas, even when arriving at the incorrect prediction.

To further investigate the different approaches the neural net took to diagnosing the entity, we tested whether one single modality was sufficient for diagnosis by training and testing only on one MR sequence. This severely diminished the algorithm's performance, with none of the four sequences yielding meaningful diagnoses alone. Compared to human raters, who heavily relied on T1CE, the neural network relies on more than one modality, and performs only slightly better than chance when evaluated on only one modality. Interestingly, when iteratively removing modalities, its performance most decreased when T2 sequences were unavailable.

While we originally incorporated an automated preprocessing step for segmenting tumor volumes, we found that our model indeed performs only slightly worse when solely run on unannotated data. This finding was further confirmed by saliency mapping, which confirmed an intuitive focus of the net toward the tumor, even when not supplied with the segmentation mask.

Finally, by combining manual diagnosis with our CNN, we demonstrated 95% accuracy, suggesting our algorithm to be well suited, not to replace expert clinicians, but to aid and provide them with an objective rating in especially challenging cases.

In conclusion, in this pilot study, we developed a novel CNN based on the DenseNet169 architecture for preoperative diagnosis of PCNSL and GBM. Our approach does not necessitate laborious preprocessing or manual annotation and can be readily used on standard MRI data. We compared the algorithm's performance against six experienced physicians and could show non-inferiority. By integrating manual rating and this automated workflow we showed 95% preoperative accuracy when detecting GBM/PCNSL. Future studies and prospective evaluation of our methodology may lead to developing a clinically useful diagnostic tool for predicting brain-tumor entities.

Limitations of the study

Although blinded to the histopathological diagnosis, raters were aware of the subject of the study being the distinction between GBM and PCNSL. This potentially biased them toward choosing one of these two diagnoses and taking special care to evaluate PCNSL features normally dismissed. However, this ensured maximal comparability with neural network performance, which, too, only had the binary choice.

The current study has further limitations, inherent to retrospective analyses. Firstly, many patients who underwent stereotactic biopsy for PCNSL could not be included because of missing MR sequences. It is for this reason, that we restricted our analysis to the four imaging modalities incorporated previously. Studies suggest that especially diffusion sequences may aid in discerning PCNSL from GBM, however, these sequences have in the past not been routinely available, especially for patients who were referred to us from other hospitals with preexistent MR-imaging sufficient for clinical management. Further, it should be noted that we did not preselect GBMs exhibiting especially PCNSL-like characteristics, which would be especially difficult to discern. However, we selected patients who had undergone stereotactic biopsy, typically performed in cases where the diagnosis is not evident from the preoperative imaging. Thus, the GBM patients enrolled in our study were likely cases not unequivocally deemed GBM at the time of surgery.

Secondly, the external validation cohort was relatively small (n = 20), but at approximately 13% of the training dataset comparable to ratios commonly used in the field.

Several CNN-based models had previously been developed for the automated diagnosis of PCNSL, as discussed previously. While a direct comparison of these models would be ideal, due to proprietary code and data protection regulations, this was not possible in this study. We were therefore restricted to comparing the performance of our model based on the published metrics of others (Table S2). Future prospective validation or pooling of data and computational approaches from multiple centers promises to increase both the number of cases used for training, as well as evaluation, promising further improvements. Additionally, prospective verification of our method on a larger number of cases would be ultimately required before this method could be clinically applicable.

ETHICS APPROVAL

This retrospective chart review study involving human participants was in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The Human Investigation Committee (IRB) of Heidelberg University was extensively consulted and waived informed patient consent due to the retrospective character of the investigation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Patient characteristics
- METHOD DETAILS
  - Radiological assessment
  - Preprocessing for automated image computing
  - Automated tumor segmentation
  - Automated tumor classification
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109023.

ACKNOWLEDGMENTS

This publication was supported through state funds approved by the State Parliament of Baden-Württemberg for the Innovation Campus Health + Life Science Alliance Heidelberg Mannheim issued to P.V.N. The authors thank Andreas Unterberg and Karl Kiening, as well as all other colleagues of the departments for fruitful discussions. The authors are grateful to Thomas Schmidt for his help with data curation and Philipp Gustav Roth for his feedback on the graphical abstract.

AUTHOR CONTRIBUTIONS

J.-O.N., P.N., and P.V.N. conceived the study. M.M., P.V.N., and M.F. performed experiments. K.K.-J., C.B.-S., A.A.B., M.J., J.J., and J.-O.N. were the human raters. P.V.N. and M.M. analyzed data and wrote the initial draft of the manuscript. C.J. contributed data and conceptual insights. K.M.-H., S.M.K., and M.B. provided infrastructure and funding. All authors reviewed the manuscript and provided feedback.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 21, 2022
Revised: December 21, 2023
Accepted: January 22, 2024
Published: January 24, 2024

REFERENCES

1. Grochans, S., Cybulska, A.M., Simińska, D., Korbecki, J., Kojder, K., Chlubek, D., and Baranowska-Bosiacka, I. (2022). Epidemiology of Glioblastoma Multiforme–Literature Review. Cancers *14*, 2412. https://doi.org/10.3390/cancers14102412.
2. Schaff, L.R., and Mellinghoff, I.K. (2023). Glioblastoma and Other Primary Brain Malignancies in Adults: A Review. JAMA *329*, 574–587. https://doi.org/10.1001/jama.2023.0023.
3. Villano, J.L., Koshy, M., Shaikh, H., Dolecek, T.A., and McCarthy, B.J. (2011). Age, gender, and racial differences in incidence and survival in primary CNS lymphoma. Br. J. Cancer *105*, 1414–1418. https://doi.org/10.1038/bjc.2011.357.
4. Eraky, A.M., Beck, R.T., Treffy, R.W., Aaronson, D.M., and Hedayat, H. (2023). Role of Advanced MR Imaging in Diagnosis of Neurological Malignancies: Current Status and Future Perspective. J. Integr. Neurosci. *22*, 73. https://doi.org/10.31083/j.jin2203073.
5. Kunimatsu, A., Kunimatsu, N., Kamiya, K., Watadani, T., Mori, H., and Abe, O. (2018). Comparison between Glioblastoma and Primary Central Nervous System Lymphoma Using MR Image-based Texture Analysis. Magn. Reson. Med. Sci. *17*, 50–57. https://doi.org/10.2463/mrms.mp.2017-0044.
6. Neumann, J.-O., Campos, B., Younes, B., Jakobs, M., Jungk, C., Beynon, C., Deimling, A.v., Unterberg, A., and Kiening, K. (2018). Frame-based stereotactic biopsies using an intraoperative MR-scanner are as safe and effective as conventional stereotactic procedures. PLoS One *13*, e0205772. https://doi.org/10.1371/journal.pone.0205772.
7. Scheichel, F., Marhold, F., Pinggera, D., Kiesel, B., Rossmann, T., Popadic, B., Woehrer, A., Weber, M., Kitzwoegerer, M., Geissler, K., et al. (2021). Influence of preoperative corticosteroid treatment on rate of diagnostic surgeries in primary central nervous system lymphoma: a multicenter retrospective study. BMC Cancer *21*, 754. https://doi.org/10.1186/s12885-021-08515-y.
8. Nguyen, A.V., Blears, E.E., Ross, E., Lall, R.R., and Ortega-Barnett, J. (2018). Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg. Focus *45*, E5. https://doi.org/10.3171/2018.8.FOCUS18325.
9. Scheichel, F., Popadic, B., Pinggera, D., Jaskolski, D.J., Lubrano, V., Foroglou, N., Netuka, D., Iliescu, B., Novak, L., Sherif, C., et al. (2023). European survey on neurosurgical management of primary central nervous system lymphomas and preoperative corticosteroid therapy. Brain Spine *3*, 101791. https://doi.org/10.1016/j.bas.2023.101791.
10. Kshatri, S.S., and Singh, D. (2023). Convolutional Neural Network in Medical Image Analysis: A Review. Arch. Comput. Methods Eng. *30*, 2793–2810. https://doi.org/10.1007/s11831-023-09898-w.
11. Malikova, H., Koubska, E., Weichet, J., Klener, J., Rulseh, A., Liscak, R., and Vojtech, Z. (2016). Can morphological MRI differentiate between primary central nervous system lymphoma and glioblastoma? Cancer Imag. *16*, 40. https://doi.org/10.1186/s40644-016-0098-9.
12. Inoue, A., Matsumoto, S., Ohnishi, T., Miyazaki, Y., Kinnami, S., Kanno, K., Honda, T., Kurata, M., Taniwaki, M., Kusakabe, K., et al. (2023). What is the Best Preoperative Quantitative Indicator to Differentiate Primary Central Nervous System Lymphoma from Glioblastoma? World Neurosurg. *172*, e517–e523. https://doi.org/10.1016/j.wneu.2023.01.065.
13. Malikova, H. (2019). Primary central nervous system lymphoma: is whole-body CT and FDG PET/CT for initial imaging reasonable? Quant. Imag. Med. Surg. *9*, 1615–1618. https://doi.org/10.21037/qims.2019.09.06.
14. Cassinelli Petersen, G.I., Shatalov, J., Verma, T., Brim, W.R., Subramanian, H., Brackett, A., Bahar, R.C., Merkaj, S., Zeevi, T., Staib, L.H., et al. (2022). Machine Learning in Differentiating Gliomas from Primary CNS Lymphomas: A Systematic Review, Reporting Quality, and Risk of Bias Assessment. AJNR. Am. J. Neuroradiol. *43*, 526–533. https://doi.org/10.3174/ajnr.A7473.
15. Kim, Y., Cho, H.-H., Kim, S.T., Park, H., Nam, D., and Kong, D.-S. (2018). Radiomics features to distinguish glioblastoma from primary central nervous system lymphoma on multi-parametric MRI. Neuroradiology *60*, 1297–1305. https://doi.org/10.1007/s00234-018-2091-4.
16. Xia, W., Hu, B., Li, H., Geng, C., Wu, Q., Yang, L., Yin, B., Gao, X., Li, Y., and Geng, D. (2021). Multiparametric-MRI-Based Radiomics Model for Differentiating Primary Central Nervous System Lymphoma From Glioblastoma: Development and Cross-Vendor Validation. J. Magn. Reson. Imag. *53*, 242–250. https://doi.org/10.1002/jmri.27344.
17. Valova, I., Harris, C., Mai, T., and Gueorguieva, N. (2020). Optimization of Convolutional Neural Networks for Imbalanced Set Classification. Procedia Comput. Sci. *176*, 660–669. https://doi.org/10.1016/j.procs.2020.09.038.
18. McAvoy, M., Prieto, P.C., Kaczmarzyk, J.R., Fernández, I.S., McNulty, J., Smith, T., Yu, K.-H., Gormley, W.B., and Arnaout, O. (2021). Classification of glioblastoma versus primary central nervous system lymphoma using convolutional neural networks. Sci. Rep. *11*, 15219. https://doi.org/10.1038/s41598-021-94733-0.
19. Yamashita, K., Yoshiura, T., Arimura, H., Mihara, F., Noguchi, T., Hiwatashi, A., Togao, O., Yamashita, Y., Shono, T., Kumazawa, S., et al. (2008). Performance Evaluation of Radiologists with Artificial Neural Network for Differential Diagnosis of Intra-Axial Cerebral Tumors on MR Images. AJNR. Am. J. Neuroradiol. *29*, 1153–1158. https://doi.org/10.3174/ajnr.A1037.
20. Park, J.E., Kim, H.S., Lee, J., Cheong, E.-N., Shin, I., Ahn, S.S., and Shim, W.H. (2020). Deep-learned time-signal intensity pattern analysis using an autoencoder captures magnetic resonance perfusion heterogeneity for brain tumor differentiation. Sci. Rep. *10*,

21485. https://doi.org/10.1038/s41598-020-78485-x.

21. Tariciotti, L., Caccavella, V.M., Fiore, G., Schisano, L., Carrabba, G., Borsa, S., Giordano, M., Palmisciano, P., Remoli, G., Remore, L.G., et al. (2022). A Deep Learning Model for Preoperative Differentiation of Glioblastoma, Brain Metastasis and Primary Central Nervous System Lymphoma: A Pilot Study. Front. Oncol. *12*, 816638. https://doi.org/10.3389/fonc.2022.816638.
22. Smith, S.M. (2002). Fast robust automated brain extraction. Hum. Brain Mapp. *17*, 143–155. https://doi.org/10.1002/hbm.10062.
23. Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage *17*, 825–841. https://doi.org/10.1006/nimg.2002.1132.
24. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., and Smith, S.M. (2012). Neuroimage *62*, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.
25. Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. Med. Image Anal. *5*, 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6.
26. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al. (2021). The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.02314.
27. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al. (2022). MONAI: An open-source framework for deep learning in healthcare. Preprint at arXiv. https://doi.org/10.48550/arXiv:2211.02701.
28. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
29. Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q. (2018). Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
30. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. Hum. Brain Mapp. *40*, 4952–4964. https://doi.org/10.1002/hbm.24750.
31. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods *18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z.
32. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241.
33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2019). Automatic differentiation in PyTorch.
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. *15*, 1929–1958.
35. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. (2020). Captum: A unified and generic model interpretability library for PyTorch. Preprint at arXiv. https://doi.org/10.48550/arXiv:2009.07896.
36. Springenberg, J.T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for Simplicity (The All Convolutional Net).
37. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In International Conference on Machine Learning, pp. 3319–3328.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Code developed in this study	This paper	https://doi.org/10.6084/m9.figshare.24986142
Automated Predictions of tumor entities	This paper	https://doi.org/10.6084/m9.figshare.24986142
Manual predictions of tumor entities	This paper	https://doi.org/10.6084/m9.figshare.25006841
Software and algorithms		
FSL BET	S.M. Smith[22]	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET
FMRIBS's Linear Image Registration Tool	Jenkinson et al.[23–25]	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT
nnUnet	Isensee et al.[26]	https://github.com/MIC-DKFZ/nnUNet
ResNet	Medical Open Network for Artificial Intelligence (MONAI)[27–29]	https://monai.io/
DenseNet	Medical Open Network for Artificial Intelligence (MONAI)[27–29]	https://monai.io/
pytorch	The Linux Foundation	https://pytorch.org/
Python	Python Software Foundation	https://www.python.org
GraphPad Prism	GraphPad Software LLC, Boston/MA, USA	https://www.graphpad.com/
Adobe Illustrator V25.4.1	Adobe Inc. San Jose, CA, USA	https://www.adobe.com/
Adobe Photoshop V25.0	Adobe Inc. San Jose, CA, USA	https://www.adobe.com/
Sankeymatic	Steve Bogart	https://sankeymatic.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Paul Naser (paul.naser@med.uni-heidelberg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data have been deposited at Figshare and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited on Figshare and is publicly available as of the date of publication. The DOIs are listed in the key resources table. The code can be additionally be obtained via https://github.com/MIC-DKFZ/iScience_GlioLymph_classification.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request (paul.naser@med.uni-heidelberg.de).

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patient characteristics

We retrospectively reviewed the patient records of our department at Heidelberg University Hospital between 01/2008 and 08/2023 and identified patients with the surgical code for stereotactic biopsy. Pathology results were reviewed for all cases, and patients diagnosed with PCNSL were identified. A similarly sized group of GBM patients from the same database was extracted. Only those patients with complete (T1, T1-CE, T2, FLAIR) high-resolution MR-imaging datasets no older than one week before stereotactic surgery were available and were included in this study. In total, 158 patients (76 male, 82 female) were enrolled. At the time of surgery, patients were aged 65.6 $\pm$ 12.5 years, with no significant differences between the groups (Figure S1). Numerically, males were slightly more prevalent in the GBM cohort (45M:34F) compared with the PCNSL cohort (39M:39M), however, this difference was not statistically significant (p=0.69; Chi-square). From the entire cohort of GBM/PCNSL patients, twenty patients (10 GBM / 10 PCNSL) were randomly chosen as the test set.

METHOD DETAILS

Radiological assessment

Radiological assessment was performed by six experienced physicians from our institution (M.J.; C.B-S.; A.A.B; J-O.N., K.K.-J. & J.J.). Volunteers were blinded to the diagnosis of the patients and were asked to diagnose the lesions based on the four imaging modalities. Further, general confidence in the radiographic diagnosis of PCNSL and experience in cranial MRI were asked. For each patient, participants completed a 12-item questionnaire ranking the importance of imaging modalities, lesion size and location, lesion attributes (contrast-enhancing tumor, non-enhancing tumor, edema, and necrosis) as well as their confidence in the diagnosis on a percentage scale. Raters assessed the items first only accessing T1, T1CE, T2 and FLAIR sequences. After finalizing their diagnosis, raters were given access to all available imaging modalities. Raters could change their diagnosis and were further asked to list the imaging sequences most important to their diagnosis beyond the aforementioned four.

Preprocessing for automated image computing

Preprocessing consisted of skull-stripping using FSL BET from the T1 images,[22] followed by utilization of the HD-BET algorithm for brain extraction of the remaining modalities.[30] Subsequently, images were registered using FMRIBS's Linear Image Registration Tool, first registering the FLAIR image to the MNI152 template, followed by coregistration of the remaining modalities to the FLAIR sequence.[23–25]

Automated tumor segmentation

For automated segmentation, a nnUnet was trained on the BraTS2021 dataset.[26,31,32] Masks were generated for peritumoral Edema (ED), Non-Enhancing Tumor (NET) and Enhancing Tumor (ET). All segmentation results were manually vetted to ensure optimal data preparation.

Automated tumor classification

The automated classification algorithm was trained with the four imaging modalities and the corresponding masks. Demographic data such as age or sex were not included in either training or test data. To classify the lesion entities, the ResNet and DenseNet architectures from Medical Open Network for Artificial Intelligence (MONAI) were initially tested.[27–29] Binary Cross-Entropy with Logits loss (BCEL) from pytorch was chosen to optimize the parameters and enable an accurate classification.[33] To assess the model's confidence, calibration, and ability to detect out-of-distribution inputs uncertainty and confidence in the automated predictions, Monte Carlo Dropout and ensemble prediction were implemented.[34] Saliency maps were computed using integrated gradients and guided backpropagation algorithms.[35–37] Overlap between saliency maps and tumor segmentation maps was calculated, and mean pixel intensities normalized to the total size of the image compared between correctly and wrongly classified patients.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data were submitted to descriptive and statistical analysis using GraphPad Prism (version 10.1.0; GraphPad Software LLC, Boston/MA, USA). The comparison of two individual groups relied on two-sided Student t-tests. When comparing multiple groups, ordinary one-way ANOVA was performed using Šídák's testing for multiple post-hoc comparisons, if applicable. Differences in standard deviation were assessed using Brown-Forsythe testing. In all cases, a p-value <0.05 was considered statistically significant and denoted by an asterisk in the figures. Additional details about the statistical analysis are provided in the figure legends. Absolute values are provided as mean $\pm$ standard deviation (s.d.).

For figure design, primary imaging data were exported via Pytorch; bar graphs, box plots, and other data illustrations as vector graphics from GraphPad Prism and arranged in Adobe Illustrator V25.4.1 and Adobe Photoshop V25.0 (Adobe Inc. San Jose, CA, USA). Sankeymatic was used to create the Sankey charts.