JOURNAL OF MACHINE LEARNING AND APPLICATIONS
Jan - Jun 2026, VOL. 2, ISSUE 1, pp. 23-29
https://doi.org/10.61577/jmla.2026.100004

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

Data provenance, technical heterogeneity, and source confounding in a composite public brain MRI dataset

Ahed J Alkhatib

Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science & Technology, Jordan (retired)

ABSTRACT

Background: MRI brain collections are often used to benchmark artificial intelligence; however, these derivatives may introduce signals from the source and preprocessing that correlate with diagnosis labels.

Objectives: To assess the data provenance, technical heterogeneity, and potential source-related confounding factors in a composite public brain MRI image dataset. It also seeks to examine whether tumor-class labels are influenced by non-anatomical image characteristics and how they threaten the validity of downstream AI model evaluations.

Methods: A file-level reproducibility audit of a composite Kaggle archive containing 6,892 JPEG MRI images labeled glioma, meningioma, pituitary tumor, or no tumor was performed. Images were assessed for clarity, exact clone, source indicators, color mode, spatial dimensions, predefined split, imaging plane, and pre-augmentation. Associations with labels were quantified using the chi-square test and Cramer's V. A transparent metadata-only rule was fitted to the predefined training partition and tested without reading pixels from the images.

Results: There were no byte-identical duplicates, and all images were readable. The archive contained a total of 5,611 images derived from BRISC, 1,281 legacies, 430 classes, 3,957 RGB files, 2,935 Gy files, and 171 explicitly augmented meningioma images. The resolution (512 × 512), mode of color (Cramer's V=0.582), source (Cramer's V=0.352), and pre-augmentation variable whose Cramers V were significantly associated (all $p<0.001$). In the untouched BRISC test subset, a rule using only color mode and resolution achieved 37.9% (95% CI 34.8-41.1%) accuracy and 45.2% balanced accuracy. For comparison, the majority baseline achieved 30.7% accuracy and 25.0% balanced-accuracy. Incorporating the source into the consolidated clean cohort enhanced the accuracy to 50.0%.

Conclusion: Technical metadata contained many label details. Therefore, random splitting at the image level or uncritical merging can inflate the apparent performance of the model. Clinical interpretation necessitates source-aware evaluation, patient- or series-level partitioning, metadata-only control, and external testing.

KEYWORDS

Brain tumor; Magnetic Resonance Imaging; Dataset audit; Source confounding; Shortcut learning

ARTICLE HISTORY

Received 11 May 2026;
Accepted 08 June 2026;
Published 15 June 2026

Introduction

Magnetic resonance imaging (MRI) plays a crucial role in intracranial tumor characterization of great quality and consistency without ionizing radiation [1]. The emergence of public datasets has expedited research on automating tumor segmentation and classification, including through the multicenter BraTS benchmarking framework and more recent collections comprising images with four different classes: glioma, meningioma, pituitary tumors, and non-tumorous [2,3]. These resources are important for developing new methods. However, the scientific value of these resources depends on the validity of the labels used, transparency of the provenance of the approaches, and evaluation partitions reflecting the intended use [3].

Brain-tumor classification studies using public contrast-enhanced T1-weighted MRI images have reported strong results using handcrafted features, convolutional neural networks, and transfer learning [4–10]. The same source images are often redistributed through multiple repositories and later recombined into new collections [9]. Consequently, the number of files in a derivative archive may not equal the number of independent patients, examinations, or acquisition sequences [10]. If related slices, transformed copies, or source-specific image formats occur in both the development and test partitions, the apparent accuracy can overstate genuine tumor recognition [10].

This concern is not hypothetical. Medical imaging models can learn the scanner identity, hospital workflow, demographic attributes, borders, text markers, and preprocessing artifacts [11–18]. Zech et al., demonstrated variable pneumonia model generalization across hospitals, while Badgeley et al., showed that hip fracture predictions could be driven largely by patient and healthcare process variables embedded in radiographs [11,19]. DeGrave et al., further demonstrated that models may use dataset-specific shortcuts rather than pathology, even when conventional discrimination appears strong [12]. These studies establish that technical and institutional signatures can act as surrogates for the disease labels.

The BRISC dataset is a new expert-reviewed dataset containing 6000 contrast-enhanced T1-weighted MRI images organized into predefined training and test partitions with segmentation masks [3]. The report also acknowledges the derivation from various public sources, the absence of systematic acquisition metadata, and the inability to ensure total subject-level independence. The archive file uploaded and examined here is a derivative and pool compilation, not the whole original BRISC release: BRISC-pattern file names are pooled along

***Correspondence:** Dr. Ahed J Alkhatib, Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science & Technology, Jordan (retired). e-mail: ajalkhatib@just.edu.jo

© 2026 The Author(s). Published by Reseapro Journals. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

with old images and class-specific augmented files.

Therefore, the objective of this study was not to claim clinical tumor classification. Instead, we investigated whether diagnostic labels in the composite archive were associated with non-anatomical technical characteristics that a model could exploit. We conducted a reproducible file-level audit, quantified the association effect sizes, evaluated a metadata-only control on preserved test partitions, and developed a defensible cohort structure for future image-model experiments. The reporting was informed by the CLAIM 2024 and TRIPOD+AI principles [20,21].

Methods

Study design and data source

This was a secondary computational audit of a publicly distributed brain MRI archive downloaded from Kaggle [22]. The archive included one root directory of four folders: Glioma, meningioma, non-tumor and pituitary. Labels from directory placement were used; there was no independent histopathologic or radiologic re-adjudication. They audited the archive as it was received before any model-specific enhancement.

File-level integrity and provenance audit

The ZIP file was tested for compression defects. Pillow read every non-directory entry that was verified to be a JPEG. The audit described for each file the archive path, filename, folder label, source family (inferred from the filename convention), predefined train/test indicator, BRISC image index, filename-derived class, color mode, anatomical plane, sequence indicator, width, height, aspect ratio, size, and SHA-256 checksum. The equality of SHA-256 digests means the same thing as byte-identical duplication. The present analysis did not exclude patient-level or near-duplicate data because patient identifiers and original series identifiers were not available.

The BRISC-derived subset was assigned filenames beginning with brisc2025_. Filenames that began with Tr- and Te- were designated as legacy subsets and thus retained their training or testing names. The files with aug in their names were found to be pre-augmented. The encoded BRISC class token label and folder label were consistent.

Cohort definitions

The main cohort consisted of 5,611 BRISC-derived images in the uploaded archive. The original partition used for testing was preserved. A deterministic 15% validation subset was selected only from the original BRISC training partition, which was stratified jointly by the diagnostic class and anatomical plane. The other BRISC training images were used to form the development set. The secondary clean cohort was composed of all 6721 images remaining after excluding the 171 explicitly pre-augmented meningioma files. The retained images were legacy images for the sensitivity analysis.

Technical variables and statistical analysis

This study designed a complete statistical validation protocol targeted at the validity of medical image diagnostic labels. First, we defined four control technical variables at the full-sample level: source family, JPEG color mode, spatial size category, and pre-enhancement. Then, for the BRISC source subset, we added two more evaluation dimensions: anatomical plane, and the original training/testing split. We adopted Pearson's chi-squared analysis for contingency tables, used Cramer's V as the effect size, and set $p<0.05$ as the significance threshold for two-tailed tests. Given the large sample size, we centered effect sizes in result interpretation, rather than relying exclusively on p-values.

Metadata-only negative-control classifier

A clear lookup rule was used as a negative control. In the training partition, the most frequent diagnosis was assigned to each unique technical metadata pattern. Subsequently, the metadata patterns of the test images were used for their classification without any decoding or analysis of the pixels of these images. The BRISC subset used the JPEG color mode and 512 × 512 status as overriding rules. The relaxation rule included the source family in the clean combined cohort. Performance evaluation was performed using accuracy, balanced accuracy, macro F1 score, and recall for each class. From the fixed test predictions, 2,000 nonparametric bootstrap resamples were drawn to obtain 95% confidence intervals (CIs). The majority class prediction served as the baseline comparator.

Preprocessing feasibility assessment

To develop future image models, primary cohort images were decoded, converted to three-channel RGB format, and resized to 224 × 224 pixels. Resizing uses aspect ratio-preserving padding with black margins. No geometric stretching occurred. The mean and standard deviation of the channel were calculated exclusively from developmental images. No test image aided the preprocessing estimation. The inability to finish GPU training in the audited environment led to the deep learning performance not being reported.

Ethics and reproducibility

Publicly available de-identified images were used in this study to train the model. No patient was contacted, nor were their identifiable clinical records accessed. To promote reproducibility, we preserved the per-image manifest, checksums, cohort flags, analysis scripts, statistical output, and figure code.

Results

Dataset composition and integrity

The collection consisted of 6892 JPEG images occupying 158,774,662 bytes (158.7MB) uncompressed. Opening of all files has been done successfully. In all filenames having the BRISC pattern, no label-token conflicts were found. Also, no byte-identical duplicate groups were encountered. Images with glioma were 1,734 (25.2%) images, meningioma 1,625 (23.6%), no tumor 1,870 (27.1%), and pituitary tumor 1,663 (24.1%) (Table 1).

Table 1. Diagnostic composition of the uploaded MRI archive.

Diagnostic label	Images, N	Percentage (%)
Glioma	1,734	25.2%
Meningioma	1,625	23.6%
No tumor	1,870	27.1%
Pituitary tumor	1,663	24.1%
Total	6,892	100.0%

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004

Figure 1 shows representative examples from the primary BRISC-derived subset.

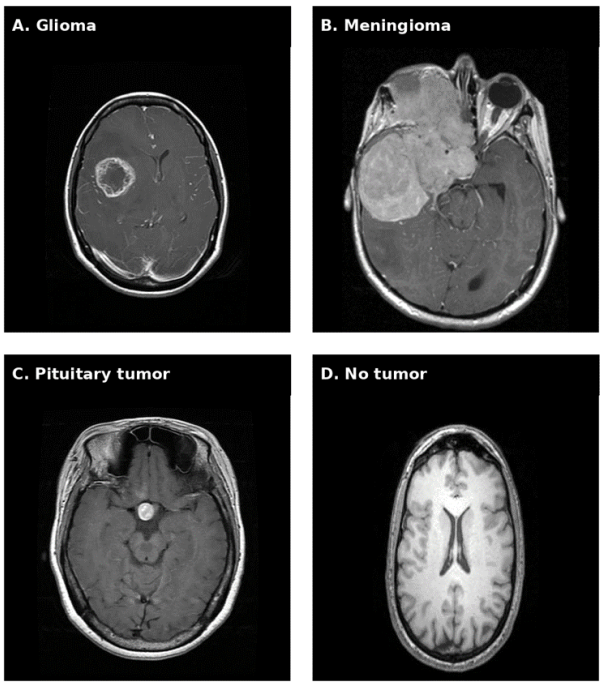


Figure 1. Representative MRI images from the primary BRISC-derived subset. Panels show examples of the four diagnostic categories included in the composite archive: (A) glioma, (B) meningioma, (C) pituitary tumor, and (D) no tumor.

Provenance and technical heterogeneity

According to filename conventions, 5,611 images from the BRISC and 1,281 legacy images were identified. The contribution to legacy strongly depended on class: it made up 37.2% of the no-tumor images. But only 0.8% of pituitary images. The archive had 430 width-height combinations. In total, 3957 images were stored in RGB format, and 2935 were stored in grey color. It must be noted that even RGB encoding, which is a three-channel store, does not necessarily suggest the presence of color MRI information in these images and may simply be a repository-specific conversion. Five thousand one hundred sixty images were of 512x512 px. Augmentation was only done to 171 meningioma images (Table 2).

Table 2. Technical characteristics by diagnostic label.

Label	Total	BRISC	Legacy	RGB, N (%)	512 x 512, N (%)	Pre-augmented
Glioma	1,734	1,344	390	374 (21.6)	1,670 (96.3)	0
Meningioma	1,625	1,444	181	945 (58.2)	1,494 (91.9)	171
No tumor	1,870	1,174	696	1,858 (99.4)	382 (20.4)	0
Pituitary	1,663	1,649	14	780 (46.9)	1,610 (96.8)	0
Total	6,892	5,611	1,281	3,957 (57.4)	5,156 (74.8)	171

According to Figure 2, the composite archive consists of 6,892 MRI images which are classified fairly evenly into four diagnosis categories. Nonetheless, the contribution of image sources varies significantly by class. The no tumor category has the largest legacy source contribution, while the pituitary tumor category is almost entirely from BRISC. The imbalance of sources indicates the potential for source-label confounding. As such, downstream models may be able to distinguish diagnostic classes in part based on technical or provenance-related image characteristics rather than tumor morphology alone.

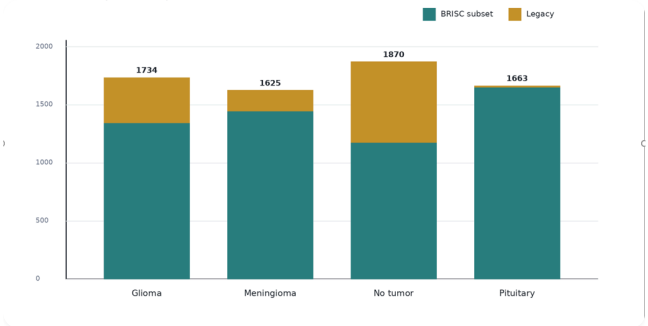


Figure 2. Class composition separated by BRISC-derived and legacy source families. Source contribution differed substantially among diagnostic labels.

Figure 3 displays a wide-ranging variation of technical image characteristics across diagnostic labels. Almost all of the no-tumor images were encoded in RGB, whereas this encoding was far less common in glioma and pituitary images. Likewise, most of the images in the tumor class were at 512 × 512 pixels whereas, the no-tumor class hardly had such resolution. The findings imply that the diagnostic labels are associated with file-level technical signatures. It is a risk that downstream AI models will learn the source or preprocessing artifacts instead of tumor morphology.

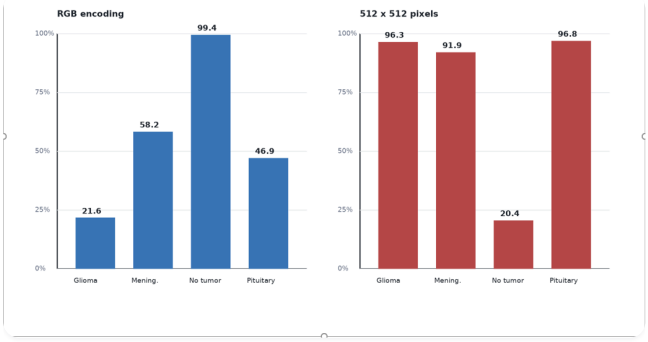


Figure 3. Percentage of images encoded as RGB and percentage stored at 512 x 512 pixels by diagnostic label. The no-tumor class had a markedly different technical profile.

Associations between labels and technical variables

All four key technical variables were linked to diagnosis. Result from spatial-size category give the largest effect (chi-square=404040.4; Cramer's V=0.766), followed with JPEG color mode (chi-square=2332.3; V=0.582), source family (chi-square=852.1; V=0.352), pre-augmentation (chi-square=568.4; V=0.287); and all p<0.001. From the BRISC subset, the associations with anatomical plane and original split were statistically significant but small (V=0.044 and V=0.081, respectively) (Table 3).

Metadata-only negative control

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004

Table 3. Associations between diagnostic label and technical characteristics.

Cohort	Variable	chi-square	df	p value	Cramer's V	N
All images	Source family	852.1	3	<0.001	0.352	6,892
All images	JPEG color mode	2332.3	3	<0.001	0.582	6,892
All images	512 x 512 status	4040.4	3	<0.001	0.766	6,892
All images	Pre-augmentation	568.4	3	<0.001	0.287	6,892
BRISC subset	Anatomical plane	21.4	6	0.002	0.044	5,611
BRISC subset	Original split	36.6	3	<0.001	0.081	5,611

Table 4 shows the rule using just JPEG color mode and 512 x 512 status achieved 37.9% accuracy (95% CI 34.8-41.1%), 45.2% balanced accuracy (95% CI 42.8-47.5%) and a macro F1-score of 34.6% on the pristine BRISC test subset. The baseline for the majority class achieved a 30.7% accuracy and a 25.0% balanced accuracy. The rule was able to identify 74.6% glioma and 83.1% no-tumor images but no meningioma images, indicating technical predictability but clinically unacceptable discrimination.

In the joint analysis of the secondary clean, the test accuracy was improved to 50.0% (95% CI 47.4-52.7%) after adding source family, as opposed to the 28.4% majority baseline. The balanced accuracy was once again at 45.6% which means the classes are not valid diagnostic. Therefore, the apparent overall accuracy was improved with the addition of source information.

Table 4. Performance of metadata-only negative-control rules.

Cohort	Technical inputs	Train/test, n	Accuracy, %	Balanced accuracy, %	Macro F1, %	Majority accuracy, %
BRISC subset	Mode + size	4,635/976	37.9	45.2	34.6	30.7
Clean combined	Source + mode + size	5,372/1,349	50.0	45.6	39.0	28.4

Source confounding and technical heterogeneity in public brain MRI datasets for artificial intelligence

The classification experiment conducted in the BRISC-derived test subset as a test in the metadata-only scenario can be seen in Figure 4. The predictions were generated solely based on JPEG color mode and image resolution, and not on pixel information. Although this restriction was serious, technical metadata alone correctly classified 370 of 976 test images. Images of glioma and no-tumor were most often predicted correctly whereas meningioma was never predicted correctly and pituitary likely assigned as glioma. The discovery demonstrated that diagnostic labels were partially encoded in non-anatomical technical properties. It showed possible presence of the source-related confounding in the composite.



Figure 4. Confusion matrix for the metadata-only rule evaluated on the untouched BRISC test subset. No pixel values or anatomical features were analyzed; meningioma recall was zero.

Leakage-controlled primary partition

Table 5 shows high-level summary of the partitioning of primary BRISC. In total, 5,611 images, comprising 3,938 development images, 697 validation images, and 976 untouched test images were included. All categories of diagnosis were represented in each partition, although totals by class. The largest class consisted of pituitary tumor images, and the smallest class of no-tumor images. The breakdown allows for designing and specifying models that are leakage attentive.

Table 5. Primary BRISC subset partitions.

Class	Development	Validation	Test	Total
Glioma	928	164	252	1,344
Meningioma	982	174	288	1,444
No tumor	882	156	136	1,174
Pituitary	1,146	203	300	1,649
Total	3,938	697	976	5,611

Leakage-controlled partitioning of the primary BRISC-derived subset

The primary BRISC-derived subset is partitioned by leakage control as shown in figure [5]. The development, validation, and untouched test partitions were formed from the overall dataset across the four diagnostic categories. The biggest

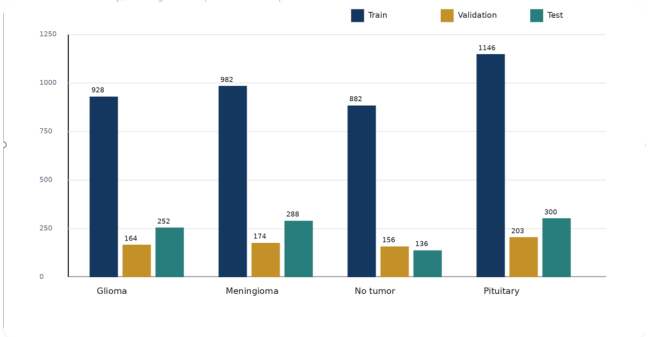


Figure 5. Class distribution across the development, validation, and untouched test partitions of the primary BRISC subset.

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004

category was images containing a pituitary tumor while the smallest was no-tumor. To prevent data leakage which will lead to an over-optimistic estimate of performance, and to keep the modelling and validation procedures transparent for later estimation of the performance of the final model, the original test partition was kept aside.

Discussion

Principal findings

The examination showed the composite archive was technically intact but methodologically heterogeneous. While it is reassuring that there are no corrupt or byte-identical files, this does not guarantee patient independence or protection against near-identical images. Most importantly, diagnostic labels were largely associated with resolution, color coding, source family, and pre-augmentation. A rule that never inspected MRI pixels outperformed the majority baseline on held-out test partitions. This research says technical metadata contained label information and could facilitate shortcut learning.

Do not take the outcome as a useful metadata classifier. It did not work at all for meningioma and macro F1 performance was weak. It was a diagnostic exercise to show that class labels could be partially recovered from file properties. In a standard image-classification experiment, a neural network may be able to observe the same properties through the canvas size, padding pattern, compression signature, intensity distribution, and processing specific to the source. A high-capacity model may combine these cues with anatomy and produce impressive internal performance whose clinical meaning is uncertain.

Relationship to BRISC and published brain-tumor classification

The official BRISC data descriptor reports 6,000 expert-reviewed contrast-enhanced T1-weighted images, including 5,000 training and 1,000 testing images. The uploaded compilation contained only 5,611 BRISC-pattern images and then added 1,281 legacy files. It therefore should not be described as the complete BRISC release. The BRISC paper also documents that its images were collated from earlier public datasets, that comprehensive scanner and acquisition metadata were unavailable, and that complete subject-level separation could not be guaranteed [3]. Those limitations are important even before derivative repackaging.

Published brain-tumor studies have reported high classification performance using public MRI collections and transfer learning [4–10]. Cheng et al., reported that tumor-region augmentation and partitioning improved three-class classification performance, while later studies used pre-trained convolutional networks and transfer learning to address limited sample size [5–8] [4–8]. More recent architectures have also reported very high internal accuracy on Kaggle-derived or public MRI datasets [9,10]. Such results demonstrate algorithmic potential, but comparisons are difficult when repositories share source images, use different split units, or include preprocessed and augmented copies. The present audit does not invalidate those studies or the BRISC labels; it shows why each derivative archive requires its own provenance and leakage assessment.

Consistency with shortcut-learning evidence

The observed technical-label associations are consistent with the evidence across medical imaging. According to Zech et al., pneumonia-detection performance differs by hospital system [19]. Badgeley et al., demonstrated the ability of scanner and process features to dominate hip-fracture prediction [11]. The work of DeGrave et al., revealed that diverse text markers, borders, positioning, and repository differences enabled COVID-19 models to opt for shortcuts instead of pathology [12]. According to Gichoya et al., deep learning can retrieve patients' race from their medical images even when it is not obvious to clinicians [13]. According to Oakden-Rayner et al., hidden stratification can hide major subgroup failure under acceptable global metrics [17].

More recent studies with formalization of shortcut test and automated domain-shift detection [14, 15]. According to Hill et al., radiographic models can be misleading in predicting behavior. Their study showed that models can easily lose face validity [16]. MRI also gets fooled by spurious contours, as Tinauer et al., showed that skull-stripping makes models used as shortcuts in AD classification more inaccurate [18]. The authors utilize these findings to suggest that a basic operational property showed the file attributes to be label informative.

Implications for future model development

A defensible continuation of this project should treat dataset design as part of model design. At minimum, future experiments should:

- Use the official source release and preserve its predefined evaluation partition rather than randomly redistributing a derivative mixture.
- Split by patient or imaging series when identifiers exist; when they do not, explicitly state that independence cannot be guaranteed.
- Remove pre-augmented images before partitioning and apply augmentation dynamically to development images only.
- Normalize dimensions and channel representation while recognizing that normalization may not eliminate compression, intensity, or source signatures.
- Report source-stratified, plane-stratified, and per-class performance with confidence intervals, calibration, and complete confusion matrices.
- Include metadata-only and border-only negative controls to test whether non-anatomical cues remain predictive.
- Perform external testing on an independently acquired institutional cohort before suggesting diagnostic utility.

Strengths and limitations

Among the strengths noted are the complete enumeration of the contents of the archive, verification of the readability of images, cryptographic checking of exact duplicates, explicit separation of source families, preservation of original test indicators and estimates of effect size. Furthermore, there was a negative-control classifier that did not have access to pixels from the image. The audit also produced a row-level manifest and deterministic cohort definitions, permitting tracking of every reported count to a file.

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004

Interpretation is constrained by several limitations. In the first instance, folder labels were not independently corroborated against pathology. Second, patient, scanner, institution and series identifiers were not available; hence subject-level independence could not be verified. SHA-256 analysis will detect identical files byte-for-byte but not near-identical transformed ones. Source attribution was inferred from filenames was fourth. The archive is a snapshot and may not represent a later version from Kaggle. Sixth, it is not to be seen as a diagnostic model as the metadata-only rule was purposefully simple. Ultimately, though deep-learning training and external clinical testing were not completed, we make no claims about tumor-recognition accuracy or clinical utility.

Conclusion

The combined public brain MRI archive contained large technical heterogeneity that was systematically associated with diagnostic labels. Label information was found in resolution, JPEG color mode, source family, and pre-augmentation status. A rule that used just technical metadata outscored the majority baseline on untouched test partitions. As a result, a significant division of an image at random would end up rewarding recognition of repository and preprocessing signatures rather than tumor biology. It is possible to use the dataset for further methodological research as long as the respect for provenance, control of augmented files, source awareness of partitions, reporting of negative controls, and limitation of conclusions to internal benchmarking all takes place. Clinical claims need evaluation data that is obtained externally and is independent of the patient.

Data and Code Availability

The Kaggle dataset page provided for this project includes the image archive to be audited. Analysis resulted in generation of per-image manifest, audit summary, preprocessing report, statistical tables, figures and reproducible python scripts.

Funding Statement

The author declares that this study received no specific external funding.

AI Disclosure

Manuscript preparation involved utilizing Generative AI-assisted tools for language polishing and organizing and drafting. The author reviewed and approved all scientific content, analysis, interpretation, reference, and final wording, and all such responsibility is of the author.

Author Contributions

Ahed J. Alkhatib conceptualized the idea, developed the methodology, curated and analyzed the data, conducted the investigation, created the visual presentation, and wrote the original draft, as well as reviewing and editing.

Ethics Declaration

The brain MRI images used in this study were previously published and de-identified. No contact was made with participants or individuals, and no records were accessed. As a result, no new informed consent was obtained, and the research was considered to not require institutional ethical approval, under the journal and author's institutional policies.

Conflict of Interest

The author declares no conflict of interest.

References

1. Musthafa N, Memon QA, Masud MM. Advancing Brain Tumor Analysis: Current Trends, Key Challenges, and Perspectives in Deep Learning-Based Brain MRI Tumor Diagnosis. Eng. 2025;6(5):82. https://doi.org/10.3390/eng6050082
2. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al.. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging. 2015;34(10):1993–2024. https://doi.org/10.1109/tmi.2014.2377694
3. Fateh A, Rezvani Y, Moayedi S, Rezvani S, Fateh F, Fateh M, et al.. BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. Scientific Data. 2026;13(1):361. https://doi.org/10.1038/s41597-026-06753-y
4. Cheng J, Huang W, Cao S, Yang R, Yang W, Yun Z, et al.. Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition. PLOS ONE. 2015;10(10):e0140381. https://doi.org/10.1371/journal.pone.0140381
5. Deepak S, Ameer PM. Brain tumor classification using deep CNN features via transfer learning. Computers in Biology and Medicine. 2019;111:103345. https://doi.org/10.1016/j.compbiomed.2019.103345
6. Sultan HH, Salem NM, Al-Atabany W. Multi-Classification of Brain Tumor Images Using Deep Neural Network. IEEE Access. 2019;7:69215–69225. https://doi.org/10.1109/access.2019.2919122
7. Swati ZNK, Zhao Q, Kabir M, Ali F, Ali Z, Ahmed S, et al.. Brain tumor classification for MR images using transfer learning and fine-tuning. Computerized Medical Imaging and Graphics. 2019;75:34–46. https://doi.org/10.1016/j.compmedimag.2019.05.001
8. Ali Khan H, Jue W, Mushtaq M, Umer Mushtaq M. Brain tumor classification in MRI image using convolutional neural network. Mathematical Biosciences and Engineering. 2020;17(5):6203–6216. https://doi.org/10.3934/mbe.2020328
9. Reddy KR, Rajesh KNVPS, Dhuli R, Kumar VR. BrainCDNet: a concatenated deep neural network for the detection of brain tumors from MRI images. Frontiers in Human Neuroscience. 2024;18:1405586. https://doi.org/10.3389/fnhum.2024.1405586
10. Rasheed M, Jaffar MA, Akram A, Rashid J, Alshalali TAN, Irshad A, et al.. Improved brain tumor classification through DenseNet121 based transfer learning. Discover Oncology. 2025;16(1):1645. https://doi.org/10.1007/s12672-025-03501-3
11. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al.. Deep learning predicts hip fracture using confounding patient and healthcare variables. npj Digital Medicine. 2019;2(1):31. https://doi.org/10.1038/s41746-019-0105-1
12. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence. 2021;3(7):610-619. https://doi.org/10.1038/s42256-021-00338-7
13. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L, et al.. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health. 2022;4(6):e406–e414. https://doi.org/10.1016/s2589-7500(22)00063-2
14. Brown A, Tomasev N, Freyberg J, Liu Y, Karthikesalingam A, Schrouff J. Detecting shortcut learning for fair medical AI using shortcut testing. Nature Communications. 2023;14(1):4314. https://doi.org/10.1038/s41467-023-39902-7
15. Guo X, Gichoya JW, Trivedi H, Purkayastha S, Banerjee I. MedShift: Automated Identification of Shift Data for Medical Image Dataset Curation. IEEE Journal of Biomedical and Health Informatics. 2023;27(8):3936–3947. https://doi.org/10.1109/jbhi.2023.3275104
16. Hill BG, Koback FL, Schilling PL. The risk of shortcutting in deep learning algorithms for medical imaging research. Scientific Reports. 2024;14(1):29224. https://doi.org/10.1038/s41598-024-79838-6
17. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the ACM Conference on Health, Inference, and Learning. 2020:151–159. https://doi.org/10.1145/3368555.3384468
18. Tinauer C, Sackl M, Stollberger R, Schmidt R, Ropele S, Langkammer C.

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004

Skull-stripping induces shortcut learning in MRI-based Alzheimer's disease classification. Insights into Imaging. 2025;16(1):283. https://doi.org/10.1186/s13244-025-02158-4

19. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. . 2018. https://doi.org/10.48550/arxiv.1807.00431.
20. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al.. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378
21. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al.. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology: Artificial Intelligence. 2024;6(4):e240300. https://doi.org/10.1148/ryai.240300
22. Kanduri R. Brain tumour dataset [dataset on the Internet]. .

© 2026 Reseapro Journals
https://doi.org/10.61577/jmla.2026.100004