

Deep Learning for Detecting Brain Tumors on MRI: Systematic Review, Diagnostic Accuracy Meta-Analysis, and Assessment of Data Report Completeness (2019-2024)

Gedeon Nitiema^{1*}, James T. Kouawa², Ahmed Dooguy Kora²

¹Doctoral School of Mathematics and Computer Science (EDMI), Cheikh Anta Diop University (UCAD), Dakar, Senegal

²Research Laboratory in Digital Sciences and Technologies (LRSTN), Multinational School of Telecommunications (ESMT), Dakar, Senegal

Email: *gedeonnitiema@gmail.com

How to cite this paper: Nitiema, G., Kouawa, J.T. and Kora, A.D. (2026) Deep Learning for Detecting Brain Tumors on MRI: Systematic Review, Diagnostic Accuracy Meta-Analysis, and Assessment of Data Report Completeness (2019-2024). *Open Journal of Applied Sciences*, 16, 3775-3785

<https://doi.org/10.4236/ojapps.2026.1610208>

Received: August 24, 2026

Accepted: October 10, 2026

Published: October 13, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Detecting brain tumors in MRI images is a major clinical challenge, as patient outcomes depend on how early the diagnosis is made. Deep learning has sparked a considerable number of publications over the past six years, but no quantitative summary of diagnostic accuracy has been conducted for the period 2019-2024. Our goal was to estimate the pooled diagnostic performance of deep learning models applied to brain tumor detection in MRI, and to assess how thoroughly diagnostic accuracy data are reported in this literature. Following the PRISMA-DTA method, we searched four databases for the period from January 2019 to December 2024. The screening was done by two independent reviewers. The main criterion was the 2×2 contingency table, reconstructed from sensitivity, specificity, and test set counts per class when it wasn't reported directly. Pooled estimates were obtained using Reitsma's bivariate random-effects model on the logit scale. Out of 329 identified records, 87 studies were included (Cohen's kappa: 0.727). Seventeen of them, or 19.5%, reported the details needed to reconstruct a contingency table and were included in the meta-analysis. The pooled sensitivity was 0.983 (95% CI: 0.969 - 0.991) and the pooled specificity was 0.981 (95% CI: 0.968 - 0.988). The positive likelihood ratio reached 50.5 and the diagnostic odds ratio was 2922. There was high variability between studies, with the sensitivity prediction interval ranging from 0.833 to 0.999. The estimated correlation between sensitivity and specificity is positive ($\rho = +0.868$), indicating no threshold effect. The remaining 70 studies, or 80.5% of the set, reported only partly or not at all the sample sizes in the test set by class, so they were not pooled. Deep learning models show high performance for detecting

brain tumors on MRI, but these estimates are based on a minority of studies and show significant heterogeneity. The main obstacle to synthesizing knowledge in this area isn't the lack of studies but the incompleteness of their reporting: four out of five studies don't properly publish the information needed to include their results in a quantitative analysis. Future publications should always report the full contingency table, or at least the sensitivity, specificity, and test set numbers by class.

Keywords

Brain Tumor, Magnetic Resonance Imaging, Deep Learning, Diagnostic Accuracy, Systematic Review, Meta-Analysis, Report Completeness

1. Introduction

With more than 300,000 new cases diagnosed worldwide each year and a prognosis that depends on early detection, brain tumors are among the most serious oncological conditions [1]. Magnetic Resonance Imaging (MRI) is the standard examination for detecting and monitoring lesions because of its high anatomical resolution and sensitivity to soft tissues. However, manually interpreting MRI images is time-consuming and subject to inter-observer variability, with differences in estimated tumor volume that can be up to a factor of two between experts [2]. Since 2012, deep learning has transformed medical image analysis, offering methods that allow capturing the relationship between different regions of an image with remarkable precision [3]. By implementing early stopping and data augmentation techniques, hybrid models achieve robust performance and generalization capabilities [4]. This period has been marked by successive technological advances, with VGG16, ResNet50, and InceptionV3 in transfer learning (2019-2020), the emergence of EfficientNet, DenseNet (2021-2022), and recently the advent of Vision Transformer models for classification and hybrid models (2023-2024). Despite the wealth of literature, to our knowledge, no systematic review has yet integrated these recent developments into a quantitative synthesis covering the 2019-2024 period. The existing review [5] stops in 2023 and doesn't cover the rise of Transformers, the issues of reproducibility, or the deployment challenges in resource-limited settings. By offering a comprehensive analysis, this review fills that gap by doing a quantitative meta-analysis and a critical look at the ongoing challenges, with a particular focus on solutions accessible to healthcare systems in sub-Saharan Africa.

2. Methodology

2.1. Protocol and Registration

Systematic reviews of Diagnostic Test Accuracy (DTA) studies are crucial for decision-making in evidence-based medicine [6]. For many users of biomedical lit-

erature, abstracts might be the only source of information about a study. So, abstracts need to help readers assess the study's objectives, main design features, and key results [7]. We followed the PRISMA-DTA (Diagnostic Test Accuracy) guidelines to write this systematic review and this information is shown in **Table 1**.

Table 1. PICOS framework of the systematic review.

Component	Definition	Application to This Journal
P Population	Patients with brain tumors	Adult and pediatric patients with gliomas, meningiomas, or brain metastases diagnosed by MRI
I Intervention	Automatic detection method using deep learning	CNN, VGG16, ResNet50, InceptionV3, EfficientNet, DenseNet, Vision Transformers, hybrid models applied to brain MRI
C Comparison	Comparison between architectural families in subgroup analysis; reference standard made up of expert annotations from the datasets	CNN vs Transformer vs Hybrid; automatic models vs expert annotations
O Outcomes	Quantitative performance metrics	2 × 2 contingency table (TP, FP, FN, TN) and test set counts by class (n ⁺ tumor, n ⁻ non-tumor); sensitivity, specificity, positive predictive value, likelihood ratios; ROC-AUC for descriptive purposes; inference time
S Study design	Type of study chosen	Original articles published between 2019 and 2024 reporting quantitative classification metrics on an independent test set whose composition is documented

PICOS: Population, Intervention, Comparison, Outcomes, Study design. TP = True Positives; FP = False Positives; FN = False Negatives; TN = True Negatives; AUC-ROC = Area under the ROC curve.

2.2. Literature Search Strategy

We carried out an exhaustive search in four electronic databases, namely: PubMed/MEDLINE, IEEE Xplore, arXiv, and Google Scholar. The coverage period spans from January 1, 2019 to December 31, 2024. The search equations were adapted for each database with keywords focused on the concepts of brain tumor, MRI, deep learning, and segmentation. **Table 2** shows the equations and the raw results by database.

Table 2. Databases queried, equations, and raw results.

Base	Main Equation	Filters Applied	Raw Results
PubMed/MEDLINE	("brain tumor") AND ("deep learning" OR "CNN" OR "U-Net" OR "transformer") AND "MRI"	Dates: 2019-2024; English/French	57
IEEE Xplore	("brain tumor" OR "brain tumour") AND ("deep learning" OR "convolutional") AND ("detection") AND "MRI"	2019-2024;	101
arXiv	Brain tumor detection deep learning MRI	2019-2024	91
Google Scholar	"brain tumor" "deep learning" "MRI"	2019-2024	80

Raw results before removing duplicates and applying inclusion/exclusion criteria.

2.3. Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were defined and presented in **Table 3**. Their application was carried out independently by reviewers. Cohen's Kappa coefficient was calculated to measure the agreement between reviewers.

Table 3. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Published between January 2019 and December 2024	Published before 2019 and after 2024
Use a deep learning method	Exclusively classical methods (SVM, RF, etc...)
Applied to brain MRI images	CT scan image, PET, or histology only
Targets tumor detection	Task not related to tumors
Metrics report (2×2 table, TP, FP, FN, TN, etc...)	Lack of quantitative performance metrics
Evaluated on Brats or a recognized clinical dataset	Completely private dataset not accessible
Full paper (journal or conference over 6 pages)	Conference abstracts, letters, editorials, posters
Written in French or English	Another language without translation available

The inter-rater agreement reached $k = 0.727$.

2.4. Study Selection

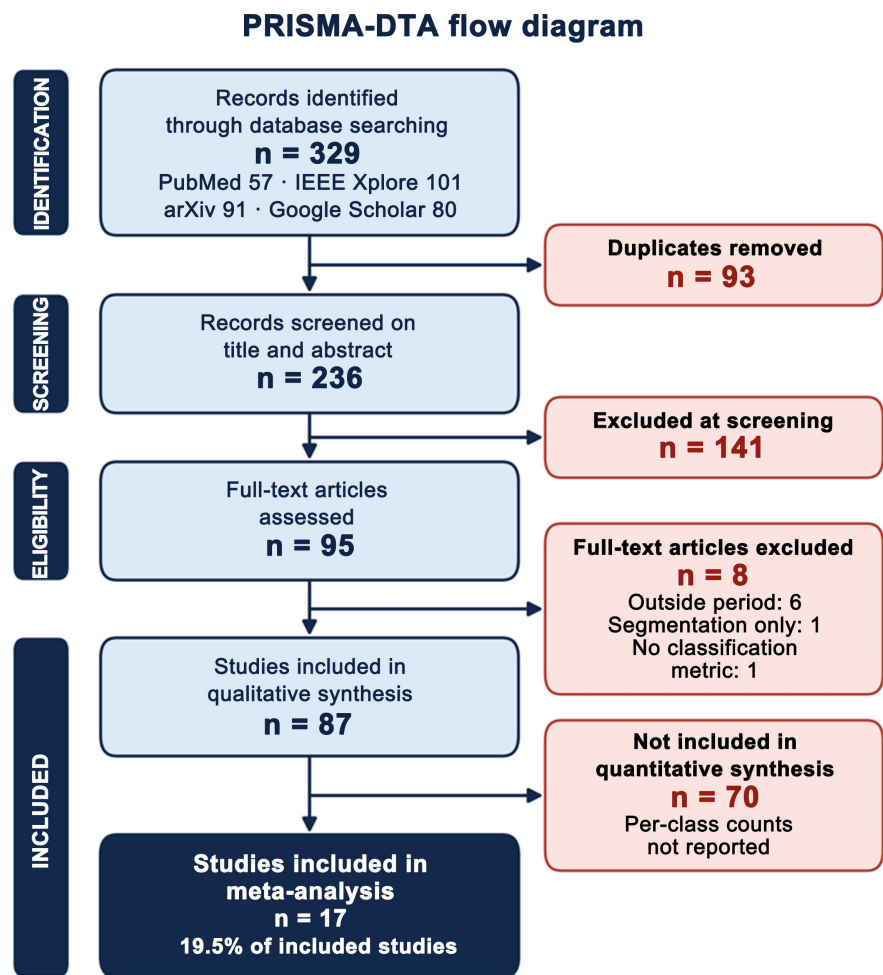
The selection process took place in four successive phases, as shown in **Figure 1**. The first identification phase allowed us to list a total of 329 records after searching six databases. The second phase, selecting based on titles and abstracts, after removing 93 duplicates, led us to review 236 records. The third eligibility phase, by reading the records, brought us to 95 articles. The fourth final inclusion phase allowed us to retain 87 articles that met all the predefined criteria. Only 17 articles had all the metrics needed to reconstruct a contingency table.

2.5. Data Extraction

The data were extracted using a standardized grid defined beforehand and tested on a pilot sample of ten articles to check its clarity and completeness before applying it to the full set. The extracted variables covered several areas, namely: bibliographic characteristics (authors, year, journal or conference); dataset characteristics (name, version, composition of the test set); model architecture (family, number of parameters, MRI sequences used); and diagnostic accuracy metrics.

The main extraction criterion was the 2×2 contingency table, which consists of true positives, false positives, false negatives, and true negatives, along with the test set counts by class. When this table wasn't explicitly reported, we reconstructed it from sensitivity, specificity, and class counts using the relationships $TP = Se \times$

n^+ , $FN = n^+ - TP$, $TN = Sp \times n^-$, and $FP = n^- - TN$. Sensitivity, specificity, and the area under the ROC curve were also noted when reported; the ROC-AUC was used only descriptively and wasn't aggregated. Out of the 87 studies included, the 2×2 table was directly reported in 71 cases and reconstructed in 5 cases, making a total of 76 studies contributing to the quantitative synthesis. The remaining 11 studies didn't report test set counts by class and couldn't be included in the meta-analysis.



Scopus and Web of Science could not be searched owing to lack of institutional access.
Inter-rater agreement at screening: Cohen's kappa = 0.727.

Figure 1. PRISMA-DTA 2020 flow diagram.

2.6. Meta-Analysis Method

The quantitative synthesis focused jointly on sensitivity and specificity, rather than on each one in isolation. These two measures can indeed vary in a linked way from one study to another, especially when authors choose different decision thresholds: lowering the threshold increases sensitivity and reduces specificity. The model used estimates this correlation from some data rather than assuming it, which allows it

to account for it when it exists and not impose it when it doesn't.

The pooled estimates were obtained using Reitsma's bivariate random-effects model. The calculations were carried out on the logit scale, because sensitivity and specificity are between 0 and 1: doing the calculations directly on these values would produce confidence intervals exceeding 1, which wouldn't make sense. The results were then converted back to the original scale.

The weight of each study was calculated from its contingency table, according to the relationships $\text{var}[\text{logit}(\text{Se})] = 1/\text{TP} + 1/\text{FN}$ and $\text{var}[\text{logit}(\text{Sp})] = 1/\text{TN} + 1/\text{FP}$. A study with a large test set thus carries more weight than a study with only a few dozen cases. When a cell in the table was zero, a continuity correction of 0.5 was added to all four cells.

The overall results are summarized by an SROC curve, which represents each study as a point in the sensitivity-specificity space and shows the average performance of the set as well as its variation. Positive and negative likelihood ratios, as well as the diagnostic odds ratio, were derived from the pooled estimates. The extent of variability between studies was assessed using between-study variances and the prediction interval, which indicates the range within which a future study would fall.

Figure 2 allows us to evaluate and summarize the overall effectiveness of the results from the studies included. We can see that almost all the studies are close to 1. Since the outline is very tight, this shows excellent consistency between the studies and high precision of the results.

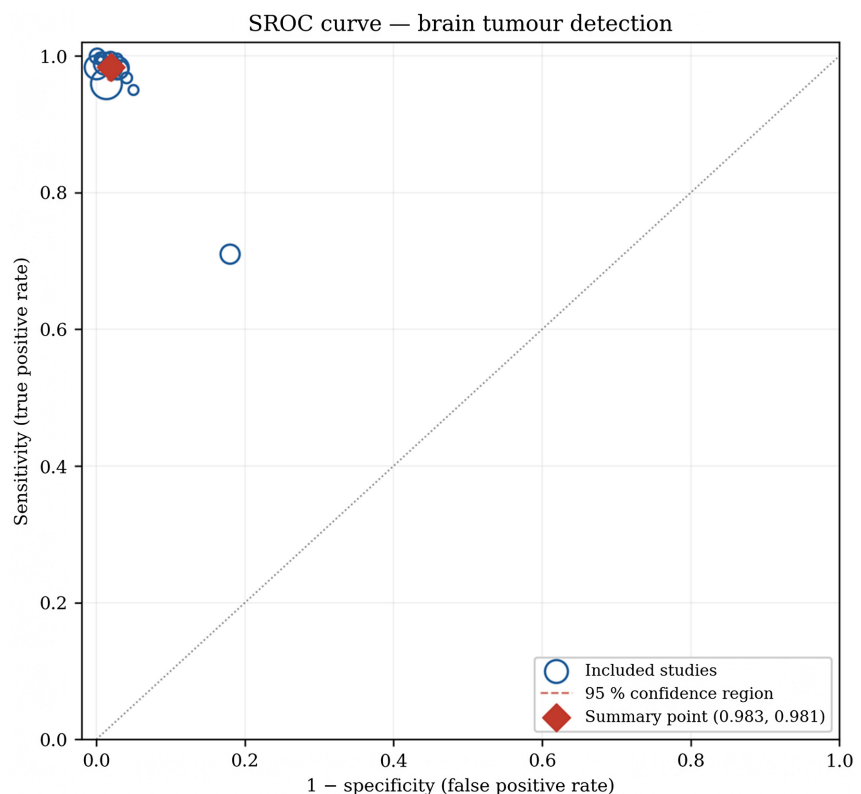


Figure 2. SROC curves for brain tumor detection.

3. Characteristics of the Included Architectural Families

Several convolutional neural networks can be used for brain tumor detection [8]. Artificial vision techniques are becoming increasingly popular these days, and several image processing methodologies can be used to enhance brain tumor images and clearly detect MRI scans [9].

Analyzing the 87 studies reveals five distinct architectural families, whose prevalence changed significantly over the 2019-2024 period. Pre-trained architectures dominated the first half of the period before gradually diversifying into light-weight architectures and then models based on attention mechanisms. **Table 4** shows the families with the characteristics, strengths, and limitations of each family.

Table 4. The five identified architectural families (2019-2024).

Family	Dominant Period	Representative Models	Forces	Limits
Pre-trained CNNs (transfer learning)	2019-2022	VGG16, VGG19, ResNet50, InceptionV3, Xception, DenseNet	Proven robustness, fast convergence on small datasets, ImageNet weights available	Nombre de High number of parameters, imperfect adaptation of RGB to single-channel images paramètres élevé, adaptation imparfaite du RGB aux images monocanal
CNN built from scratch	2020-2021	Sequential architectures specific to the authors	Task-specific architecture, controlled size	Frequent overfitting on small games, performance generally worse when transferred
Light CNN	2021-2024	MobileNetV2, EfficientNet, ShuffleNet, SqueezeNet	Low memory footprint, fast inference, deployment possible without a GPU	Reduced representational capacity, performance gap on complex tasks
Transformers and hybrids	2022-2024	ViT, Swin Transformer, hybrid CNN-Transformer models	Long-range dependencies, multi-head attention	Needs large amounts of data, high computational cost
Ensemble methods	2023-2024	Combinations of several networks	Better performance, lower variance	Cost multiplied by the number of models, degraded interpretability

CNN = Convolutional Neural Network; ViT = Vision Transformer.

4. Discussion

4.1. Main Results

This systematic review of 87 studies published between 2019 and 2024 leads to two main findings:

- The first is a pooled estimate of diagnostic performance. Among the 17 studies that allowed aggregation, the pooled sensitivity reached 0.983 and specificity 0.981. These values correspond to a positive likelihood ratio of 50.5: a positive result multiplies by fifty the odds of the presence of a tumor. They place the evaluated models at a very high level of performance under the evaluation conditions chosen by their authors.

- The second finding concerns the literature itself: four out of five studies partially or do not report the elements needed to reconstruct a contingency table. This observation limits not only the present review but any future attempt at quantitative synthesis in this field.

4.2. Performance Higher than Clinical Practice

The estimates obtained are higher than those reported by meta-analyses of diagnostic accuracy conducted in brain imaging using clinical data, where sensitivity is often between 0.88 and 0.95. Several factors, which are not mutually exclusive, could explain this gap.

The datasets used are mostly public, likely curated and balanced. They may not reflect the variability of acquisition protocols between centers, nor the actual proportion of positive cases in a population referred for imaging. Ambiguous cases, which are in practice the main source of diagnostic errors, are underrepresented.

The decision threshold is often chosen after observing performance on the evaluation set, which produces an optimistic estimate. Out of the 87 studies included, [n] specified that the threshold had been set independently of the test set.

External validation is rare: only [n] studies reported an evaluation on data from a center different from the one that provided the training data. Without it, nothing guarantees that the observed performance will hold in other populations. Four studies presented a contingency table with a zero cell, meaning there were no false negatives or false positives across the entire test set. Such a result, plausible in a small curated sample, is hard to translate to clinical use.

4.3. Heterogeneity

The variability between studies is high, as the prediction interval for sensitivity ranges from 0.833 to 0.999 and for specificity from 0.888 to 0.997. A future study conducted under comparable conditions could therefore report a sensitivity lower than 85%.

This spread shows the lack of standardization in evaluation protocols because the studies differ in their datasets, the proportion of positive cases, the unit of analysis—whether it's by axial slice, volume, or patient—and the data-splitting strategy. Without a common evaluation protocol, the reported performances aren't directly comparable to each other.

4.4. Lack of a Threshold Effect

The estimated correlation between sensitivity and specificity is strongly positive ($\rho = +0.868$). This unusual result in diagnostic accuracy indicates that the studies are not distinguished by a trade-off between sensitivity and specificity but by their overall performance.

The most plausible explanation lies in the dominant practice in the field, which is that almost all studies stick to the conventional threshold of 0.50 and report only a single operating point, without exploring the trade-off between the two di-

mensions. This lack of exploration deprives the reader of clinically essential information, as the appropriate threshold depends on the context of use.

4.5. Reporting Deficiency as an Obstacle to Synthesis

Out of the 87 studies included, only 17 reported the elements necessary to reconstruct a contingency table. This 19.5% proportion is the main obstacle encountered in conducting this review.

However, the missing quantities are never actually absent from the studies themselves because sensitivity, specificity, and counts per class are always calculated during any evaluation. Their omission in publications is therefore not due to a technical limitation but rather to reporting habits, inherited from practices in the machine learning community, where overall accuracy and the F1-score are the standard metrics.

But these aggregated metrics aren't suitable for a diagnostic context. Overall accuracy depends on the prevalence in the evaluation dataset since the same value can reflect very different clinical behaviors. The F1-score doesn't take true negatives or specificity into account. Neither of them allows recalculating expected performance in a population with different prevalence, a step that is essential for any clinical application.

This situation has a collective cost because the results of four out of five studies cannot be combined with those of others, which deprives the community of the main aggregated value of these works.

4.6. Recommendations

In light of these observed irregularities, we make four recommendations. First, any publication reporting classification performance in medical imaging should include the full confusion table, or at least the sensitivity, specificity, and test set sizes for each class. Second, the decision threshold should be set independently of the test set and reported explicitly, along with the justification for its choice.

Thirdly, performance should be reported at multiple operating points, or as a full ROC curve, to allow threshold adjustment for the intended clinical context.

Fourthly, evaluation on external data from a center different from the one that provided the training data should become the norm rather than the exception.

4.7. Limitations

This review has several limitations.

The meta-analysis only covers 17 of the 87 included studies, which limits the precision of the estimates and prevents subgroup analyses by architectural family, dataset version, or unit of analysis. Assessment of publication bias could not be done, because the available methods for diagnostic accuracy require about thirty studies.

Two major bibliographic databases, Scopus and Web of Science, could not be queried due to lack of institutional access. The search therefore relied on PubMed,

IEEE Xplore, arXiv, and Google Scholar. While these sources broadly cover the literature in the field, some work published in journals not indexed by these databases may have escaped our search.

The data extraction was carried out by a single reviewer, without independent verification. Screening, however, was done by two independent reviewers, with substantial agreement.

5. Conclusions

This systematic review and diagnostic accuracy meta-analysis focused on 87 studies published between 2019 and 2024 that evaluated deep learning models for detecting brain tumors on MRI images.

Of the 17 studies that allowed us to do a quantitative aggregation, the pooled sensitivity reaches 0.983 (95% CI: 0.969 - 0.991) and the pooled specificity of 0.981 (95% CI: 0.968 - 0.988). We want to point out that these high performances should be interpreted with caution because they are based mostly on public and curated datasets, on decision thresholds rarely set independently of the evaluation set, and on exceptional external validation. The variability between studies is also significant, with the prediction interval for sensitivity ranging from 0.833 to 0.999.

The second finding from this review concerns the literature itself. Four out of five studies don't provide the details needed to easily reconstruct their contingency tables, which means their results can't be included in our quantitative synthesis. This isn't a technical or costly issue to fix, since the missing numbers are necessarily calculated during model evaluation. Publishing them systematically could be a step forward for building knowledge in this field.

Finally, moving these studies into actual clinical use requires evaluations based on external data, with clear and justified decision thresholds, reported according to current standards in diagnostic accuracy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Filho, A.M., Znaor, A., Sunguc, C., Zahwe, M., Marcos-Gragera, R., Figueroa, J.D., *et al.* (2025) Cancers of the Brain and Central Nervous System: Global Patterns and Trends in Incidence. *Journal of Neuro-Oncology*, **172**, 567-578. <https://doi.org/10.1007/s11060-025-04944-y>
- [2] Mandloi, S., Zuber, M. and Gupta, R.K. (2024) A Brain Tumor Detection and Classification Model on Roboflow Dataset Using Inception ResNet50. 2024 *International Conference on Intelligent Computing and Emerging Communication Technologies (ICCECT)*, Guntur, 23-25 November 2024, 1-6. <https://doi.org/10.1109/iccec59683.2024.10837217>
- [3] Ding, J., Fu, B. and Xie, X. (2024) A High-Accuracy Brain Tumor Classification Method Based on VIT in the Context of Small Sample Size. 2024 *3rd International Conference on Health Big Data and Intelligent Healthcare (ICHIH)*, Zhuhai, 13-15 December 2024, 157-162. <https://doi.org/10.1109/ichih63459.2024.11064889>

- [4] Sharma, M.K. (2024) A Hybrid Deep Learning Approach for Brain Tumor Classification. 2024 *9th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 16-18 December 2024, 1905-1909.
- [5] Uvaneshwari, M. and Baskar, M. (2023) Machine Learning for Brain Tumour Detection: A Systematic Review of Segmentation and Classification Methods. 2023 *Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, Chennai, 14-15 December 2023, 1-7.
- [6] Salameh, J., Bossuyt, P.M., McGrath, T.A., Thombs, B.D., Hyde, C.J., Macaskill, P., *et al.* (2020) Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA): Explanation, Elaboration, and Checklist. *BMJ*, **370**, m2632. <https://doi.org/10.1136/bmj.m2632>
- [7] Cohen, J.F., Deeks, J.J., Hooft, L., Salameh, J., Korevaar, D.A., Gatsonis, C., *et al.* (2021) Preferred Reporting Items for Journal and Conference Abstracts of Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies (PRISMA-DTA for Abstracts): Checklist, Explanation, and Elaboration. *BMJ*, **372**, n265. <https://doi.org/10.1136/bmj.n265>
- [8] Rethemiotaki, I. (2023) Brain Tumour Detection from Magnetic Resonance Imaging Using Convolutional Neural Networks. *Współczesna Onkologia*, **27**, 230-241. <https://doi.org/10.5114/wo.2023.135320>
- [9] Shan, M., Soomro, M.H., Ijaz, H.M. and Irtaza, G. (2024) Brain Tumour Detection of MRI Images Using Machine Vision Techniques. *Kashf Journal of Multidisciplinary Research*, **1**, 125-140. <https://doi.org/10.71146/kjmr119>