VNUHCM Journal of Health Sciences 2026, 7(2):1033-1044
Open Access Full Text Article

Research Article

Efficient brain tumor detection in MRI scans using YOLOv11 with pretrained backbones

Thanh-Hai Le[1], Doan-Van Anh Ly[2], Si-Nguyen Phan[2], Thi-Thu-Hien Pham[3,4,*]

Use your smartphone to scan this QR code and download this article

ABSTRACT

[Background] Precise object detection within medical imaging constitutes a fundamental requirement for computer-aided diagnostic systems, particularly for the localization of brain tumors. Brain tumors present significant diagnostic challenges due to their heterogeneity in shape, size, texture, and anatomical location across varying Magnetic Resonance Imaging (MRI) modalities. While manual interpretation by clinical specialists is the gold standard, it remains labor-intensive, time-consuming, and subject to inter-observer variability. Consequently, automated deep learning systems have emerged as vital tools for rapid screening. However, balancing detection accuracy with computational efficiency remains a critical challenge for real-world clinical deployment, particularly when moving beyond simple classification to precise spatial localization.

[Objective] This study investigates the performance efficacy of integrating distinct Convolutional Neural Network (CNN) backbones into the state-of-the-art YOLOv11 architecture to optimize tumor detection in MRI scans. Our primary aim is to systematically evaluate how variations in feature extraction architecture influence the critical trade-off between localization precision and computational complexity.

[Methods] Four model variants were engineered and tested by substituting the native YOLOv11 backbone with structurally diverse pretrained networks: ResNet50, ResNeXt50, SEResNet50 (incorporating attention mechanisms), and MobileNetV3-Large (optimized for mobile architectures). These models were trained and validated on a manually annotated dataset comprising three primary tumor classes: glioma, meningioma, and pituitary tumors. Ground truth bounding boxes were strictly delineated to capture exact tumor spatial extents. Performance was rigorously quantified utilizing standard object detection metrics—Precision, Recall, mAP@0.5, and mAP@0.5:0.95—alongside an evaluation of model parameter counts to assess computational weight.

[Results] Comparative analysis reveals a distinct performance dichotomy dictated by backbone design. Deeper architectures—specifically ResNet50 and SEResNet50—yield superior feature extraction capabilities, maximizing detection accuracy and achieving the highest overall mAP scores. Conversely, MobileNetV3 demonstrates a highly favorable efficiency-performance trade-off. Despite a significantly reduced parameter count and lower computational overhead, the MobileNetV3-backed model retains competitive mAP scores, proving its capability to accurately localize abnormalities efficiently.

[Conclusion] These findings establish distinct deployment strategies tailored to clinical needs. ResNet-based backbones are optimal for precision-critical diagnostic tasks where computational resources are abundant, whereas MobileNetV3 offers a viable, lightweight alternative for implementation in resource-constrained clinical hardware or real-time preliminary screening. Future work will focus on validating these lightweight architectures across multi-center datasets to ensure broader generalizability.

Key words: Brain tumor localization, YOLOv11, Lightweight backbones, Magnetic resonance imaging

[1]*School of Computer Science and Engineering, The Saigon International University, Ho Chi Minh City 700000, Vietnam*

[2]*Department of Information Technology, FPT University, Ho Chi Minh City 700000, Vietnam*

[3]*School of Biomedical Engineering, International University, Ho Chi Minh City, Vietnam*

[4]*Vietnam National University Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam*

Correspondence

Thi-Thu-Hien Pham, School of Biomedical Engineering, International University, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

Email: ptthien@hcmiu.edu.vn

History
- Received: 15-8-2025
- Revised: 30-12-2025
- Accepted: 28-5-2026
- Published Online: 02-07-2026

DOI : https://doi.org/10.32508/vnuhcmjhs.v7i2.700

Check for updates

INTRODUCTION

Brain tumors present a major diagnostic challenge due to their significant variability in shape, size, and location across different Magnetic Resonance Imaging (MRI) modalities. While MRI provides detailed visualization of soft tissues, manual interpretation is labor-intensive and prone to inter-observer variability. Consequently, automated computer-aided diagnosis (CAD) systems based on Deep Learning (DL) have become essential for enhancing diagnostic consistency and efficiency [1,2].

In recent years, Convolutional Neural Networks (CNNs) and hybrid architectures have dominated medical image analysis. A substantial body of research has focused on tumor segmentation, aiming to delineate exact tumor boundaries. For instance, Nguyen-Tat et al.[3] recently demonstrated that in-

Cite this article : Le T, Anh Ly D, Phan S, Pham T. **Efficient brain tumor detection in MRI scans using YOLOv11 with pretrained backbones**. *VNUHCM J. Health Sci.* 2026; 7(2):1033-1044.

tegrating Contextual Transformers with 3D U-Nets significantly enhances segmentation performance by capturing intricate contextual details. However, full pixel-wise segmentation is computationally demanding. Addressing this, recent work by Nguyen-Tat et al.[4] proposed the "DoubleBlock-ViT," a lightweight framework designed to balance global contextual awareness with computational efficiency, highlighting the growing clinical need for resource-efficient models.

Despite these advancements in segmentation, there remains a critical need for object detection—a middle ground that offers precise localization (via bounding boxes) faster than segmentation but with more spatial utility than simple classification. Early approaches focused primarily on classification[5] or compared older detection frameworks. Ahsan et al.[6] benchmarked algorithms like Faster R-CNN and SSD, identifying YOLOv5 as a superior candidate for tumor detection due to its inference speed. More recently, Taha et al.[7] explored advanced iterations like YOLOv8 and YOLOv11, reporting high validation accuracy in preliminary experiments. However, these studies often rely on default backbone configurations, which may not be optimized for the specific texture and scale variations inherent in MRI data.

A key limitation in current literature is the lack of systematic evaluation regarding how different feature extraction backbones influence the trade-off between detection precision and model complexity in medical imaging. While deeper networks (e.g., ResNet) offer robust feature extraction[5], they may introduce latency unsuitable for real-time screening. Conversely, lightweight backbones (e.g., MobileNet) offer speed but potentially at the cost of sensitivity.

To address this gap, this study investigates the architectural optimization of the YOLOv11 model for brain tumor localization. Unlike previous studies that focus on classification[8] or computationally heavy segmentation[3], we propose a detection-centric approach that balances clinical accuracy with deployment efficiency. We replace the native backbone with four distinct architectures—ResNet50, ResNeXt50, SEResNet50, and MobileNetV3-Large—and evaluate their performance on a multi-class MRI dataset. This approach allows for a direct assessment of how backbone depth and design influence the identification of glioma, meningioma, and pituitary tumors in resource-constrained environments.

MATERIALS AND METHODS

Dataset

This study utilizes a comprehensive brain tumor MRI dataset obtained from Kaggle and the Hayatabad Medical Complex Peshawar, comprising images with different resolutions from multiple MRI modalities[9]. The dataset includes four categories of brain conditions: glioma, meningioma, pituitary tumor, and no tumor (healthy brain scans) with a total of 4350 images and annotation files. Each class includes multiple views such as axial, coronal, and sagittal, ensuring a diverse representation of clinical scenarios.

The original dataset did not include object detection annotations in YOLO format. To address this, all images were manually annotated using the LabelImg tool[9]. This ensured accurate and clinically reliable bounding box labels for tumor localization.

The 'No Tumor' category, which consists of healthy brain scans and originally contained full-brain annotations without specific objects, was explicitly retained in the dataset to serve as background samples. For these images, we generated empty annotation files (containing no bounding boxes), instructing the model to treat the entire image as background. This approach allows the model to learn feature representations of healthy tissue and minimizes false positive detections. Since no predefined training/validation/testing split was provided, the dataset was manually divided into train:val:test as 0.8:0.1:0.1 (1740:217:218 images) with balanced class distribution maintained across all sets as shown in Figure 1. The three randomly selected samples from each category, along with their bounding boxes, are displayed in Figure 2 within the training set.

All images were resized to 640×640 before training. Standard augmentation techniques were applied, including horizontal flipping and color jittering. This dataset setup provided a robust foundation for evaluating the effectiveness of different backbone architectures in tumor detection.

Models

Base Model Selection

YOLO (You Only Look Once) is a widely used one-stage object detector known for its balance of speed and accuracy. Initially, both YOLOv8[10] and YOLOv11[11] were selected as baseline models to evaluate their performance across different model scales—specifically, the small (s), medium (m), and large (l) variants. These experiments were conducted to determine which model provided the best balance between detection accuracy and computational cost with no pretrained weight. Figure 3 presents the operational framework that will be utilized throughout this study.

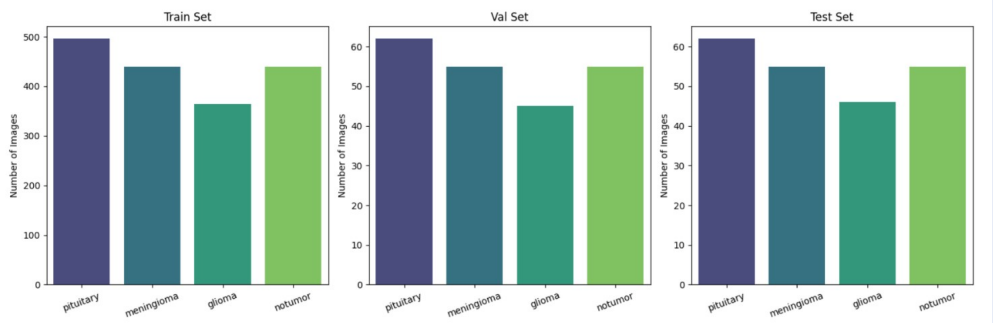

Figure 1: Distribution of each class on training, validation and testingsets. (Source: Created by the authors)

Backbone Selection and Integration

The backbone network serves as the primary feature extractor, governing the trade-off between accuracy and computational efficiency. To analyze the impact of architectural diversity on YOLO's performance, we replaced the standard backbone with four distinct architectures using the timm library[12]. The selected backbones represent a spectrum of design philosophies:

ResNet50[13]: Adopted as a robust baseline, this architecture utilizes residual connections to mitigate the vanishing gradient problem, offering an optimal balance between depth and trainability.

ResNeXt50[14]: A variant that introduces "cardinality" via grouped convolutions. It improves representational power through aggregated residual transformations without significantly increasing parameter complexity compared to ResNet.

SEResNet50[15]: This model integrates Squeeze-and-Excitation (SE) blocks into the ResNet structure. The SE mechanism explicitly models inter-channel dependencies, allowing the network to recalibrate feature maps by emphasizing informative channels and suppressing less relevant ones.

MobileNetV3[16]: Selected to evaluate efficiency on resource-constrained devices. It utilizes neural architecture search (NAS) and depth-wise separable convolutions to minimize computational cost while maintaining competitive accuracy.

All backbones were initialized with pretrained ImageNet weights to accelerate convergence. To ensure seamless integration with the YOLO detection head, we modified the architecture to extract multi-scale feature maps from three hierarchical stages (typically corresponding to stride 8, 16, and 32). These features are aligned with the Feature Pyramid Network (FPN) structure[17] of the YOLO neck, preserving the model's ability to detect objects of varying sizes. It is important to note that replacing the backbone overrides the standard YOLO scaling (s/m/l), as the parameter count is now dictated by the chosen backbone architecture.

RESULTS AND DISCUSSION

Results

Experimental setup

All experiments were conducted in a Kaggle environment, utilizing specified hardware and training hyperparameters, along with selected data augmentation techniques. Table 1 and Table 2 provide detailed information on the hardware specifications and training hyperparameters employed.

We trained the YOLOv11s and its variants with various pretrained backbone from scratch in the Kaggle environment, using a maximum of 100 epochs with an early stopping patience set to 10 to prevent overfitting. Figure 4 illustrates the training and validation performance of YOLOv11s throughout this process, highlighting the model's convergence and stability under these conditions.

Ablation study

Among YOLOv8 and YOLOv11 variants tested, YOLOv11s (small version) yielded the best performance in terms of accuracy and efficiency.

The YOLOv11 with different backbones

As detailed in Table 3, YOLOv11s achieved superior scores in three out of the four evaluation metrics compared to other variants. Consequently, it was selected as the most robust baseline for conducting further experiments with different backbone architectures, using identical training settings and evaluation metrics as outlined above.

To quantify model uncertainty and practical deployability, we analyzed the Mean $\pm$ Standard Deviation

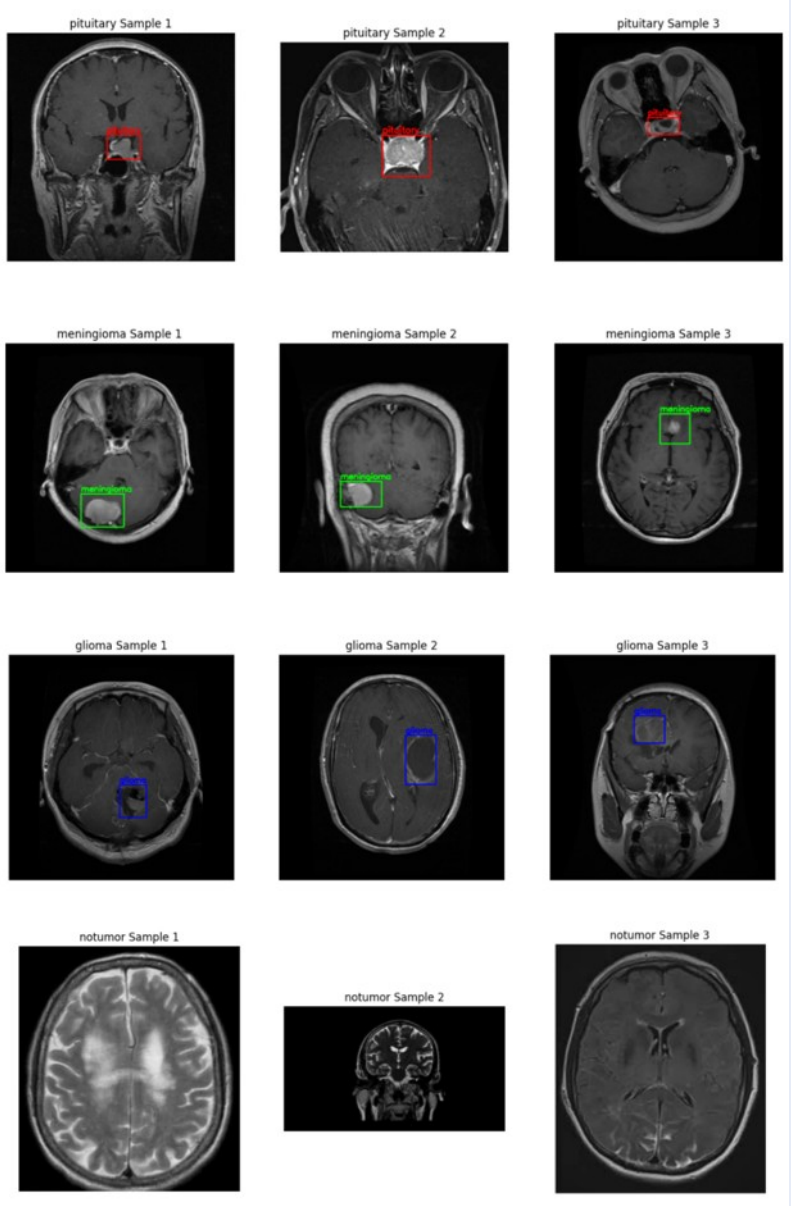

Figure 2: Three random images fromeach class with their bounding boxes. (Source: Created by theauthors)

Table 1: Kaggle experimental environment

Processor	Intel(R)Xeon(R)CPU@2.20GH
RAM	31.4 GB
Graphics card	Tesla T4 GPU
Programming language	Python 3.10.12
Deep learning framework	PyTorch 2.5.1+cu121

(Source: Created by the authors)

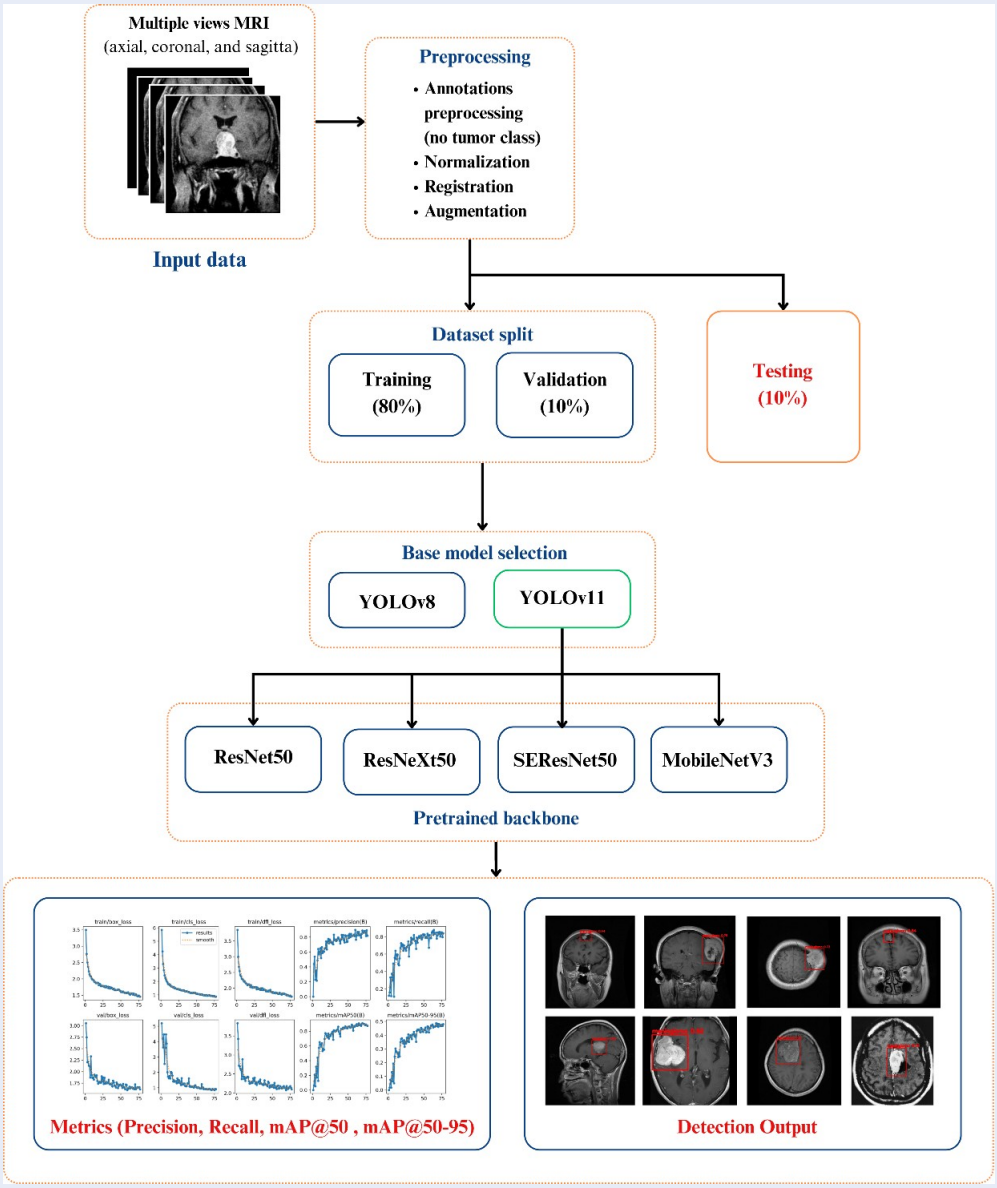

Figure 3: Research framework. (Source: Created by the authors)

Table 2: Training hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Initial learning rate	0.01	Optimizer	SGD
Batch size	16	Epoch	100
Image size	640 × 640	Momentum	0.937
Patience	10	Weight decay	0.0005

(Source: Created by the authors)

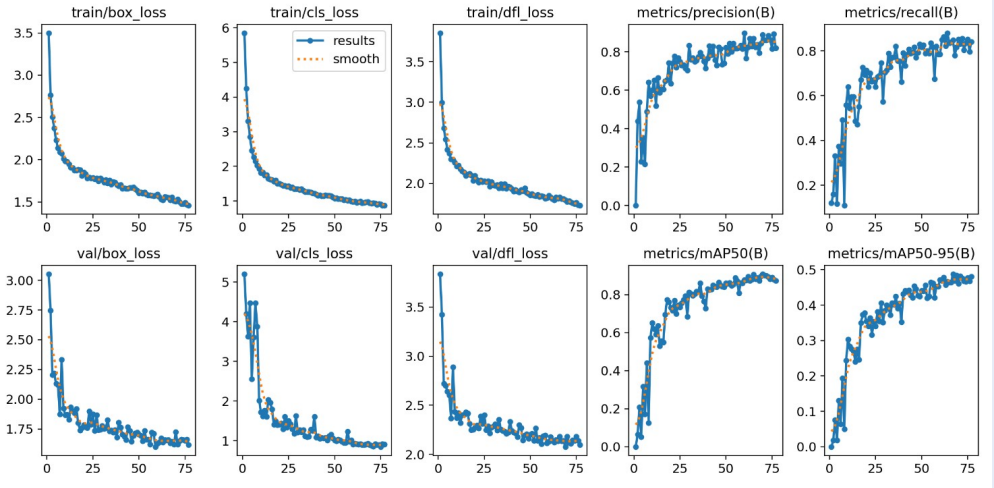

Figure 4: Training and validation process visualization of YOLOv11s model. (Source:Created by the authors)

Table 3: Performance of YOLO variants on testing set

Model	Precision	Recall	mAP@50	mAP@50-95
YOLOv8s	0.897	0.821	0.875	0.476
YOLOv8m	0.722	0.767	0.782	0.407
YOLOv8l	0.835	0.834	0.859	0.499
YOLOv11s	0.920	0.840	0.907	0.485
YOLOv11m	0.829	0.781	0.849	0.469
YOLOv11l	0.768	0.793	0.821	0.464

(Note that highest and lowest scores are highlighted in green and blue fonts, respectively.)
(Source: Created by the authors)

Table 4: Performance of YOLOv11s model and its variants on the testing set (Mean ± Standard Deviation).

Model	Parameter	Average Latency per Image (ms)	Precision	Recall	AP@50	AP@50-95
YOLOv11s	9.4M	13.80	0.8868 ± 0.2932	0.9151 ± 0.2772	0.9151 ± 0.2772	0.6335 ± 0.3051
Pretrained backbone ResNet50	32.9M	34.55	0.9128 ± 0.2702	0.9243 ± 0.2629	0.9197 ± 0.2659	0.6282 ± 0.306
Pretrained backbone ResNeXt50	32.3M	40.69	0.9052 ± 0.2714	0.9289 ± 0.2553	0.9235 ± 0.26	0.6254 ± 0.3072
Pretrained backbone SEResNet50	35.4M	41.80	0.88 ± 0.2989	0.9151 ± 0.2772	0.9128 ± 0.2786	0.6263 ± 0.3081
Pretrained backbone MobileNetV3	7.8M	18.10	0.9174 ± 0.2496	0.9427 ± 0.2305	0.9404 ± 0.2324	0.6317 ± 0.2927

(Note that highest and lowest scores are highlighted in green and blue fonts, respectively.)
(Source: Created by the authors)

(STD) of performance metrics across all test samples alongside the Average Inference Latency measured on the same hardware configuration in Table 1. As shown in Table 4, the MobileNetV3-Large backbone emerges as the optimal configuration for clinical applications. It achieved the highest mean Precision, Recall, and AP@50, offering superior stability with only a marginal latency increase (~4.3 ms) over the baseline. While the YOLOv11s baseline remains the fastest (13.80 ms) and excels in strict localization (AP@50-95: 0.634 ± 0.305), MobileNetV3 provides a far better efficiency trade-off than heavier architectures like ResNet50, which incur double the inference cost for comparable accuracy.

Discussion

This study evaluated the architectural adaptability of the YOLOv11 framework for brain tumor localization. Our empirical results demonstrate that the choice of feature extractor significantly influences the trade-off between detection precision and model complexity. The ResNet50 and ResNeXt50 backbones demonstrated robust performance, providing a superior balance of Precision, Recall, and mAP@0.5:0.95. This robustness is attributed to the deep residual connections within these networks, which enhance feature extraction capabilities. As noted by Sasaki et al.[18], gliomas are radiologically characterized by heterogeneous internal textures and infiltrative boundaries. The deeper feature hierarchies of ResNet-based architectures are necessary to preserve the high-frequency spatial information required to resolve these subtle contrasts, preventing the misclassifications between background and tumor classes observed in shallower baselines.

Conversely, the MobileNetV3-Large backbone achieved the highest precision among all tested variants. This result is particularly notable given that it operates with significantly fewer parameters than the ResNet-based models. This suggests that for distinct pathologies like meningiomas, the gross morphological features are sufficient for detection. Nawaz et al.[19] describe meningiomas as typically "globose" with well-defined margins; the relative simplicity of these shapes allows MobileNetV3 to achieve high true-positive rates—specifically for meningioma classes —without the computational overhead of dense residual blocks.

Visual inspection of inference results (Figures 6 and 7) combined with confusion matrix analysis (Figure 5) elucidates the model's reliability. While the system demonstrates precise localization for well-defined lesions (Figure 6), failure cases reveal characteristic challenges. As seen in the figure, redundant detections on heterogeneous textures (Figure 7b) suggest challenges in non-maximum suppression for complex masses. Conversely, false negatives (Figure 7a) occurred most frequently in small, low-contrast glioma sub-regions. This visual evidence supports Sasaki et al.'s[18] observation that the heterogeneous, infiltrative nature of gliomas often disrupts feature continuity in 2D slices, making them harder to distinguish from the surrounding skull base or complex parenchyma than encapsulated meningiomas.

Furthermore, it is crucial to benchmark against large-scale segmentation efforts. Zeineldin et al.[20], a winner of the BraTS 2022 challenge, achieved Dice scores approximating 0.93 using complex multimodal CNN ensembles. While such models define the upper bound of accuracy on large-scale datasets, they require heavy computational resources and full volumetric (3D) data. Our YOLOv11-MobileNetV3 configuration serves as a complementary solution: while it may not match the granular boundary definition of Zeineldin's heavy BraTS models, it provides clinically useful localization with significantly lower latency, addressing the need for rapid screening tools in 2D slice-based workflows.

Limitation

A primary limitation of this study is the reliance on data collected from a singular clinical institution. Consequently, limiting the breadth of demographic and scanner variability in our training distribution. Additionally, rigorous external validation was constrained by the nature of currently available public repositories. Since major datasets typically provide segmentation masks and classification labels rather than the bounding boxes required for object detection, direct benchmarking was not immediately feasible without extensive manual re-annotation. Future research initiatives will aim to bridge this gap by expanding data acquisition to multiple centers and converting standard segmentation datasets into detection formats, ensuring broader applicability of the proposed methodology.

CONCLUSIONS

This study benchmarked the architectural adaptability of YOLOv11 for brain tumor detection, isolating the specific trade-offs between accuracy and complexity across diverse backbones. Results indicate distinct performance profiles: ResNet50 maximized localization quality (highest mAP@0.5:0.95 of 0.511), confirming the necessity of deep residual hierarchies for granular boundary definition. Conversely,

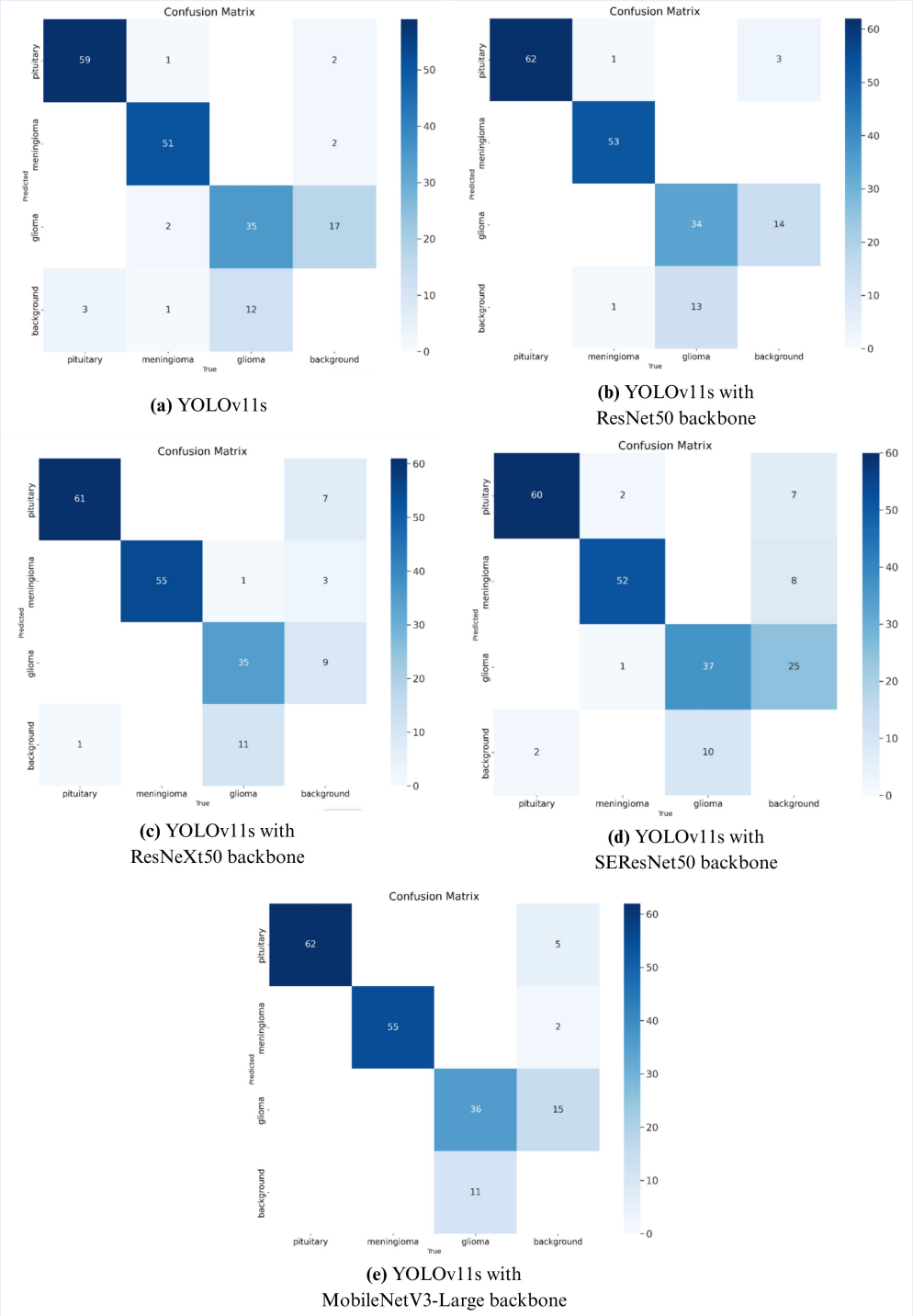

Figure 5: Performance of YOLOv11s with MobileNetV3-Large backbone for each class. (Source: Created by the authors)

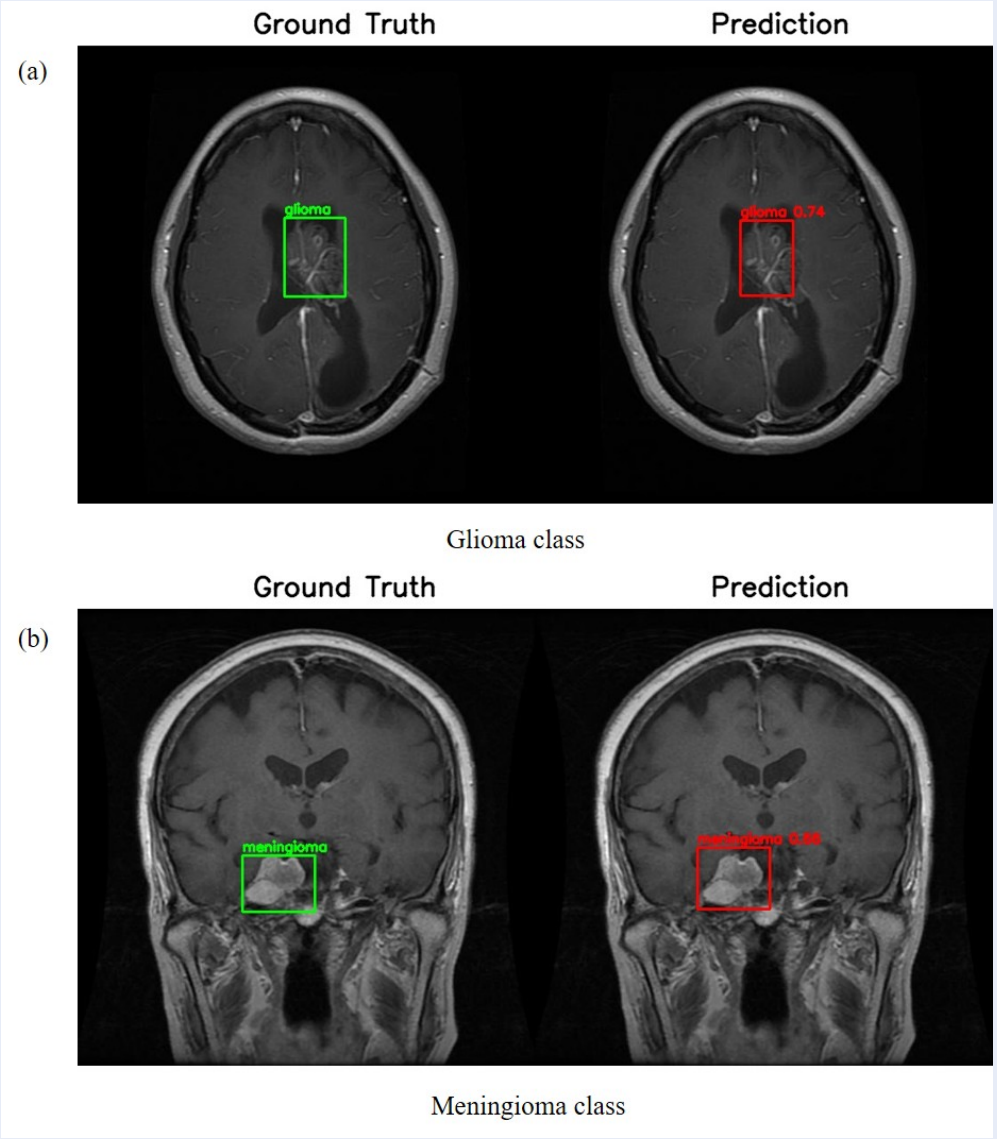

Figure 6: Correct detection result of YOLOv11s with MobileNetV3-Large backbone. (Source: Created by the authors)

MobileNetV3-Large achieved the highest raw precision (0.932) with only 7.8 million parameters. This provides technical evidence that NAS-designed backbones can retain competitive sensitivity, making them ideal for hardware-constrained environments where low latency is prioritized.

These quantitative findings directly motivate our future research directions. The observed performance gap between the robust ResNet50 and the efficient MobileNetV3 supports the investigation of knowledge distillation to transfer deep feature representations into compact student networks (<5M parameters). Furthermore, given the high precision of the MobileNet variant, we plan to conduct reader studies to quantify whether these automated proposals can effectively reduce inter-observer variability and reporting time in clinical workflows. Finally, to address the current single-center limitation, we will expand our dataset with multi-institutional cohorts and adapt public benchmarks for rigorous cross-validation, ensuring model robustness across diverse patient populations.

COMPETING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or pub-

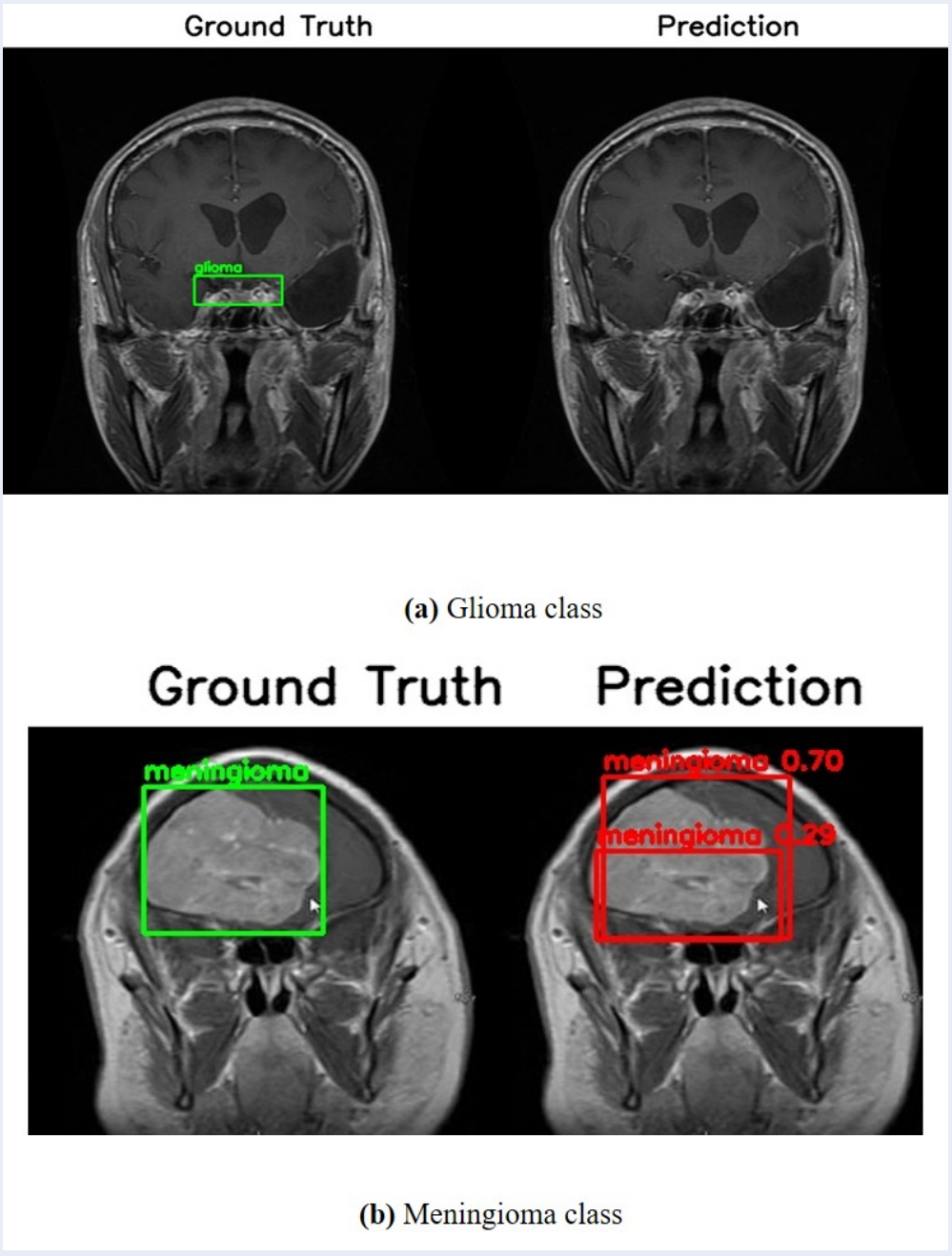

Figure 7: Incorrect detection result of YOLOv11s with MobileNetV3-Large backbone. (Source: Created by the authors)

lication of this article.

AUTHORS' CONTRIBUTIONS

Thanh-Hai Le: Conceptualization; Methodology; Validation; Software; Project administration.

Doan-Van-Anh Ly: Methodology; Validation; Software; Writing – Original Draft.

Si-Nguyen Phan: Methodology; Validation.

Thi-Thu-Hien Pham: Conceptualization; Formal analysis; Investigation; Supervision; Funding acquisition.

All authors: Reviewed, edited, and approved the final manuscript.

FUNDING

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2023-28-02.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this paper are publicly available in the Kaggle repository at https://www.kaggle.com/datasets/ammarahmed310/labeled-mri-braintumor-dataset/data[9].

REFERENCES

1. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE; 2014. p. 580–7. Available from: https://doi.org/10.1109/CVPR.2014.81.
2. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 779–88. Available from: https://doi.org/10.1109/CVPR.2016.91.
3. Nguyen-Tat TB, Nguyen TQ, Nguyen HN, Ngo VM. Enhancing brain tumor segmentation in MRI images: A hybrid approach using UNet, attention mechanisms, and transformers. Egypt Inform J. 2024;27:100528. Available from: https://doi.org/10.1016/j.eij.2024.100528.
4. Nguyen-Tat TB, Truong LQ, Truong KQ, Ngo TH. DoubleBlock-ViT: A MaxViT-based enhancement and dual-path skip connections for brain tumor segmentation in MRI scans. Computer Methods and Programs in Biomedicine. 2026;274:109165. Available from: https://doi.org/10.1016/j.cmpb.2025.109165.
5. Güler M, NamlıE. Brain tumor detection with deep learning methods' classifier optimization using medical images. Applied Sciences (Basel, Switzerland). 2024;14(2):642. Available from: https://doi.org/10.3390/app14020642.
6. Ahsan R, Shahzadi I, Najeeb F, Omer H, et al. Brain tumor detection and segmentation using deep learning. Magnetic Resonance Materials in Biology Physics and Medicine. 2025;38(1):13–22. Available from: https://doi.org/10.1007/s10334-024-01203-5.
7. Taha AM, Aly SA, Darwish MF. Detecting glioma, meningioma, and pituitary tumors and normal brain tissues based on YOLOv11 and YOLOv8 deep learning models [Preprint]. arXiv. 2025. Available from: https://arxiv.org/abs/2504.00189.
8. Amin J, Sharif M, Haldorai A, Yasmin M, Nayak RS. Brain tumor detection and classification using machine learning: a comprehensive survey. Complex Intell Syst. 2022;8(4):3161–83. Available from: https://doi.org/10.1007/s40747-021-00563-y.
9. Ahmed A, Hamza KM. "Labeled MRI brain Tumor dataset." ; 2024. Available from: https://www.kaggle.com/datasets/ammarahmed310/labeled-mri-brain-tumor-dataset.
10. Jocher G, Chaurasia A, Qiu J. "Ultralytics YOLOv8." Version 8.0.0. 2023. Available from: https://github.com/ultralytics/ultralytics.
11. Jocher G, Qiu J. "Ultralytics YOLO11." Version 11.0.0. 2024. Available from: 0.2024.URL;https://github.com/ultralytics/ultralytics.
12. Wightman R. "PyTorch Image Models." ; 2019. Available from: https://doi.org/10.5281/zenodo.4414861;https://github.com/rwightman/pytorch-image-models.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 770–778. Available from: https://doi.org/10.1109/CVPR.2016.90.
14. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE; 2017. p. 1492–1500. Available from: https://doi.org/10.1109/CVPR.2017.634.
15. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE; 2018. p. 7132–7141. Available from: https://doi.org/10.1109/CVPR.2018.00745.
16. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: and others, editor. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE; 2019. p. 1314–1324. Available from: https://doi.org/10.1109/ICCV.2019.00140.
17. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA; 2017. p. 936–944. Available from: https://doi.org/10.1109/CVPR.2017.106.
18. Kunimatsu A, Yasaka K, Akai H, Sugawara H, Kunimatsu N, Abe O. Texture Analysis in Brain Tumor MR Imaging. Magnetic Resonance in Medical Sciences : MRMS : An Official Journal of Japan Society of Magnetic Resonance in Medicine. 2022;21(1):95–109. Available from: https://doi.org/10.2463/mrms.rev.2020-0159.
19. Bhattacharjee S, Prakash D, Kim CH, Kim HC, Choi HK. Texture, Morphology, and Statistical Analysis to Differentiate Primary Brain Tumors on Two-Dimensional Magnetic Resonance Imaging Scans Using Artificial Intelligence Techniques. Healthcare Informatics Research. 2022;28(1):46–57. Available from: https://doi.org/10.4258/hir.2022.28.1.46.
20. Zeineldin RA, Karar ME, Burgert O, Mathis-Ullrich F, et al.. Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution. Cham: Springer; 2023. Available from: https://doi.org/10.1007/978-3-031-33842-7_11.

VNUHCM Journal of Health Sciences 2026, 7(2):1033-1044

Open Access Full Text Article

Bài nghiên cứu

Phát hiện khối u não một cách hiệu quả bằng ảnh cộng hưởng từ (MRI) sử dụng YOLOv11 với các backbone được huấn luyện trước

Lê Thanh Hải[1], Lý Đoàn Vân Anh[2], Phan Sĩ Nguyên[2], Phạm Thị Thu Hiền[3,4,*]

Use your smartphone to scan this QR code and download this article

[1]*Khoa Kỹ thuật và Khoa học Máy tính, Trường Đại học Tư thục Quốc tế Sài Gòn, Thành phố Hồ Chí Minh 700000, Việt Nam*

[2]*Bộ môn Công nghệ Thông tin, Trường Đại học FPT, Thành phố Hồ Chí Minh 700000, Việt Nam*

[3]*Khoa Kỹ thuật Y Sinh, Trường Đại học Quốc tế, Thành phố Hồ Chí Minh, Việt Nam*

[4]*Đại học Quốc gia Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh 700000, Việt Nam*

Liên hệ

Phạm Thị Thu Hiền, Khoa Kỹ thuật Y Sinh, Trường Đại học Quốc tế, Thành phố Hồ Chí Minh, Việt Nam

Đại học Quốc gia Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh 700000, Việt Nam

Email: ptthien@hcmiu.edu.vn

Lịch sử

- Ngày nhận: 15-8-2025
- Ngày sửa đổi: 30-12-2025
- Ngày chấp nhận: 28-5-2026
- Ngày đăng: 02-07-2026

DOI : https://doi.org/10.32508/vnuhcmj-hs.v7i1.706

Check for updates

TÓM TẮT

[Bối cảnh] Việc phát hiện đối tượng chính xác trong hình ảnh y tế là một yêu cầu cơ bản đối với các hệ thống chẩn đoán hỗ trợ bằng máy tính, đặc biệt là trong việc định vị khối u não. Các khối u não đặt ra những thách thức chẩn đoán đáng kể do tính không đồng nhất về hình dạng, kích thước, kết cấu và vị trí giải phẫu trên nhiều phương thức Chụp cộng hưởng từ (MRI) khác nhau. Mặc dù việc đọc kết quả thủ công bởi các chuyên gia lâm sàng là tiêu chuẩn vàng, quá trình này vẫn tốn nhiều công sức, thời gian và dễ bị ảnh hưởng bởi sự khác biệt giữa các bác sĩ. Do đó, các hệ thống học sâu tự động đã trở thành những công cụ quan trọng hỗ trợ sàng lọc nhanh chóng. Tuy nhiên, việc cân bằng giữa độ chính xác khi phát hiện và hiệu quả tính toán vẫn là một thách thức lớn đối với việc triển khai thực tế trên lâm sàng, đặc biệt là khi tiến xa hơn việc phân loại đơn thuần để đạt được khả năng định vị không gian chính xác.

[Mục tiêu] Nghiên cứu này khảo sát hiệu quả của việc tích hợp các mạng xương sống (backbone) là các mạng nơ-ron tích chập (CNN) khác nhau vào kiến trúc YOLOv11 tiên tiến nhất nhằm tối ưu hóa việc phát hiện khối u trong ảnh MRI. Mục tiêu chính là đánh giá một cách có hệ thống xem sự thay đổi trong kiến trúc trích xuất đặc trưng ảnh hưởng như thế nào đến sự đánh đổi quan trọng giữa độ chính xác khi định vị và độ phức tạp tính toán.

[Phương pháp] Bốn biến thể mô hình đã được lựa chọn và thử nghiệm bằng cách thay thế backbone gốc của YOLOv11 bằng các mô hình được huấn luyện trước có cấu trúc đa dạng gồm: ResNet50, ResNeXt50, SEResNet50 (tích hợp cơ chế Attention), và MobileNetV3-Large (được tối ưu hóa cho thiết bị di động). Các mô hình này được huấn luyện và kiểm định trên một tập dữ liệu được gán nhãn thủ công bao gồm ba loại khối u chính: u thần kinh đệm (glioma), u màng não (meningioma) và u tuyến yên (pituitary). Các hộp định vị (bounding boxes) được xác định rõ ràng để bao quát chính xác phạm vi – vị trí không gian của khối u. Hiệu suất được định lượng bằng các thước đo tiêu chuẩn cho bài toán phát hiện đối tượng gồm Precision, Recall, mAP@0.5 và mAP@0.5:0.95 cùng với việc đánh giá số lượng tham số của mô hình để ước lượng khối lượng tính toán.

[Kết quả] Phân tích so sánh cho thấy sự phân hóa hiệu suất rõ rệt được quyết định bởi thiết kế của backbone. Các kiến trúc sâu hơn – cụ thể là ResNet50 và SEResNet50 – mang lại khả năng trích xuất đặc trưng vượt trội, tối đa hóa độ chính xác và đạt được điểm mAP tổng thể cao nhất. Ngược lại, MobileNetV3 cho thấy sự cân bằng rất lý tưởng giữa hiệu quả và hiệu suất. Mặc dù số lượng tham số giảm đáng kể và chi phí tính toán thấp hơn, mô hình sử dụng backbone MobileNetV3 vẫn duy trì được điểm mAP cạnh tranh, chứng minh khả năng định vị chính xác các bất thường một cách hiệu quả.

[Kết luận] Những phát hiện này thiết lập các chiến lược triển khai riêng biệt phù hợp với nhu cầu lâm sàng. Các backbone dựa trên cấu trúc ResNet là lựa chọn tối ưu cho các tác vụ chẩn đoán đòi hỏi độ chính xác cao ở những hệ thống có nguồn tài nguyên tính toán dồi dào. Trong khi đó, MobileNetV3 cung cấp một giải pháp thay thế gọn nhẹ, khả thi để ứng dụng trên các thiết bị y tế bị hạn chế về phần cứng hoặc trong hệ thống sàng lọc sơ bộ theo thời gian thực. Các nghiên cứu trong tương lai sẽ tập trung vào việc kiểm định các kiến trúc gọn nhẹ này trên các tập dữ liệu đa trung tâm để đảm bảo tính khái quát hóa rộng rãi hơn.

Từ khoá: Định vị khối u não, YOLOv11, Mạng backbone gọn nhẹ, Chụp cộng hưởng từ (MRI)

Trích dẫn bài báo này: Thanh Hải L, Vân Anh L D, Sĩ Nguyên P, Thu Hiền P T. **Phát hiện khối u não một cách hiệu quả bằng ảnh cộng hưởng từ (MRI) sử dụng YOLOv11 với các backbone được huấn luyện trước**. *VNUHCM J. Health Sci.* 2026; 7(2):1033-1044.