Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

UDC 004.934:141.1 DOI: https://doi.org/10.30837/2522-9818.2026.1.017

Kristina Dostalova, Alexandra Cizmarova, Marek Klimo, Roman Yaroshevych

EXPLANATION DETECTED BRAIN TUMOURS IN MRI IMAGES USING YOLOV8 WITH LIME-BASED INTERPRETATION

Relevance. Precise identification of brain tumours in magnetic resonance imaging (MRI) is a critical task in medical image analysis. Although deep learning–based object detectors achieve high localisation accuracy, their limited transparency restricts trust and routine adoption in clinical practice, highlighting the need for explainable artificial intelligence (XAI) approaches. **Object of research.** The object of this research is the automated detection of brain tumours in MRI scans using convolutional neural network – based object detection models. **Subject of research.** The subject of the research is the integration of YOLOv8 object detection models with the Local Interpretable Model-Agnostic Explanations (LIME) method to interpret individual detection outputs in medical imaging. **Purpose.** The aim of this paper is to develop and evaluate an explainable framework for brain tumour detection in MRI images by integrating YOLOv8-based object detection with LIME-based interpretation and by quantitatively assessing the quality of the generated explanations. **Methods.** Two YOLOv8 variants (YOLOv8n and YOLOv8s) were trained and evaluated on a publicly available MRI dataset containing glioma, meningioma, and pituitary tumour classes. LIME was applied to generate superpixel-based, box-conditioned local explanations for individual detections. Detection performance was assessed using precision, recall, mAP@50, and mAP@50–95. Explanation quality was quantitatively evaluated using stability, sparsity, maximum superpixel weight, and entropy metrics. **Results.** Experimental results demonstrate that both YOLOv8 models achieve high detection performance, with YOLOv8s providing slightly improved accuracy. LIME successfully highlights image regions that most influence model decisions, and the proposed quantitative metrics confirm that the generated explanations are stable, informative, and aligned with clinically relevant tumour regions. **Conclusions.** The proposed framework provides a practical approach for combining accurate tumour localisation with interpretable and quantitatively validated explanations, supporting reliability-oriented evaluation of AI-based clinical decision-support systems.

Keywords: Explainable Artificial Intelligence (XAI); LIME; medical image analysis, brain tumour detection; MRI; YOLOv8.

Introduction

Relevance and an overview of scientific works

The use of artificial intelligence (AI) and machine learning has become increasingly widespread in modern healthcare [1]. It aligns with trends in precision and smart medicine developed under Industry 4.0 [2, 3]. This approach has the potential to reduce human error in diagnosis and treatment selection [4, 5]. This progress is particularly evident in medical image analysis, including ultrasound, X-ray, and magnetic resonance imaging (MRI) [6]. One of the key challenges in medical image processing is the accurate and timely identification of brain tumours in magnetic resonance imaging (MRI) [7]. While deep learning has significantly advanced localisation and classification performance in medical images, its limited transparency restricts trust and routine adoption in clinical practice [8, 9, 10]. Object detection models such as the YOLO family offer fast, end-to-end tumour localisation and class prediction, yet their decision paths are difficult to interrogate and validate by clinicians [11]. This motivates explainable artificial intelligence (XAI) approaches that expose which image regions and visual cues drive algorithmic decisions, thereby enabling critical scrutiny, error analysis, and safer integration into clinical workflows [10, 12, 13].

In neuro-oncology, the tension between performance and transparency is particularly acute. MRI appearances of glioma, meningioma, and pituitary tumours can overlap, and downstream clinical actions (biopsy planning, therapy selection, and follow-up strategies) depend on how much confidence care teams place in algorithmic suggestions [9]. Beyond general “interpretability”, healthcare decision support systems must satisfy broader reliability requirements: stable behaviour under minor perturbations, robustness to distributional shifts, traceability of contributing factors, and explicit treatment of uncertainty. These requirements have been widely recognised in the reliability engineering community, including for healthcare applications where human factors and organisational contexts shape outcomes, and where aleatory and epistemic uncertainties must be disentangled and managed [4, 5, 14].

Existing XAI research in medical imaging has largely focused on classification or segmentation networks (e.g., U-Net variants for BRATS-style tasks), with explanation techniques such as Grad CAM/Grad CAM++ or relevance propagation providing saliency

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

maps over the full image [1,3]. However, object detection, where the model first proposes bounding boxes and then assigns class confidences, poses distinct explanation challenges. Post hoc methods must attribute predictions to a specific detection rather than the whole image; they should reflect both localisation and classification components and be compatible with non-differentiable post-processing steps used by modern detectors (e.g., NMS). As a result, despite encouraging reports of YOLO-based medical detectors, high-quality, box-conditioned, and quantitatively assessed explanations for detection outputs remain underexplored [10, 12, 15].

To address these limitations, the present study integrates a modern one-stage detector, YOLOv8, with the LIME framework to obtain local, model-agnostic, superpixel-based explanations that are explicitly tied to individual detections (bounding boxes) in brain MRI. We evaluate two YOLOv8 variants (YOLOv8n and YOLOv8s) to study the trade-off between computational efficiency and both accuracy and interpretability, following a consistent training and evaluation protocol on a public MRI dataset comprising glioma, meningioma, and pituitary tumours. We further complement qualitative overlays with a set of quantitative explanation metrics: stability, sparsity, maximum superpixel weight, and entropy. These metrics are designed to assess how concentrated, consistent, and interpretable the explanations are in practice. This combined perspective aims to bridge performance metrics (precision, recall, mAP@50, mAP@50–95) with reliability-oriented interpretability indicators meaningful for clinical decision-making.

From a reliability engineering perspective, quantifying explanation behaviour enhances the robustness of AI systems in healthcare. Stable, sparse, and clinically aligned explanations facilitate expert verification, support effective handover between algorithm and clinician, and mitigate risks associated with spurious correlations [5, 16].

Setting objectives

The aim of this paper is to develop and evaluate an explainable framework for brain tumour detection in MRI images by integrating YOLOv8-based object detection with LIME-based interpretation and by quantitatively assessing the quality of the generated explanations. The proposed solution integrates YOLOv8-based object detection models with the LIME explainable artificial intelligence technique to improve the transparency and reliability of model predictions.

This paper makes the following contributions:

- Conditioned, model-agnostic explanations for MRI tumour detection. LIME is adapted to generate box-conditioned superpixel explanations for YOLOv8 detections in brain MRI, enabling clinicians to verify why a specific bounding box and tumour class were proposed rather than only viewing global image saliency.
- Quantitative evaluation of explanation quality. A compact set of interpretability metrics (Stability, Sparsity, MaxWeight, and Entropy) is introduced and computed to assess robustness, concentration, and distribution of importance across superpixels. These metrics operationalise reliability-oriented interpretability by linking explanation behaviour to desirable properties such as robustness and focus.
- Empirical study across model capacities. YOLOv8n and YOLOv8s are compared under identical training conditions and dataset splits to analyse how model capacity affects detection performance (precision, recall, mAP@50, mAP@50–95) and the structure and stability of generated explanations.
- Reliability-aware perspective for clinical AI. Interpretability is situated within a broader reliability engineering framework for healthcare, providing a principled rationale for adopting quantitative XAI criteria in deployment-oriented evaluations.

The methodology

Dataset

The experimental evaluation was performed using a publicly available brain tumour MRI dataset designed for automated tumour detection tasks. The dataset contains magnetic resonance images annotated with the position of pathological areas and each image is labelled according to the tumour type, including glioma, meningioma, and pituitary tumours [17].

The dataset was selected due to its compatibility with object detection frameworks and its suitability for training YOLO-based models [18]. All images were resized to a uniform resolution to ensure consistency during training and inference. The dataset was divided into training, validation, and test subsets, enabling objective evaluation of detection performance and explainability results.

This dataset provides a representative collection of MRI scans with varying tumour sizes, locations, and visual characteristics, which enables robust evaluation of both detection accuracy and model interpretability.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

YOLOv8-based brain tumour detection

Brain tumour detection was performed using the YOLOv8 [19] object detection framework, which is designed to localize objects within images by predicting bounding boxes in a single forward pass. YOLOv8 was selected for its balance between detection accuracy, computational efficiency, and flexibility for training on custom datasets.

In this study, two YOLOv8 model variants were utilised: YOLOv8n and YOLOv8s. The YOLOv8n variant is a lightweight model optimised for fast inference and reduced computational requirements, whereas YOLOv8s provides higher detection accuracy at the cost of increased model complexity. Comparing these two variants enables analysis of the trade-off between efficiency and performance in medical image applications.

Both models were trained under identical conditions to ensure a fair comparison. Training was conducted on resized MRI images with a fixed input resolution, and standard data augmentation techniques were applied to enhance model generalisation. The models were optimised using stochastic gradient-based learning and evaluated using established object detection metrics.

The detection outputs consist of bounding boxes with associated confidence scores and tumour class predictions, which serve as inputs for subsequent explainability analysis using the LIME method.

Explainable Artificial Intelligence (XAI) [12]

In recent years, a wide range of explainability methods has been proposed to address the limited transparency of deep learning models in medical imaging. Gradient-based approaches, such as Grad-CAM and Grad-CAM++, are among the most frequently used techniques. These methods visualize important regions by analyzing gradients in convolutional layers, which makes them relatively fast and easy to apply. However, their explanations are often coarse and heavily dependent on the model's internal architecture. Moreover, they are mainly designed for classification tasks and may not fully capture localized decision regions in object detection scenarios.

Another group of methods is based on relevance propagation, such as Layer-wise Relevance Propagation (LRP). These approaches aim to trace the model’s prediction back to individual input pixels, providing detailed relevance maps. While this can provide fine-grained explanations, LRP requires direct access to the model's structure and careful adaptation to specific architectures, limiting its flexibility when working with complex detection models.

Perturbation-based methods, such as SHAP and LIME, explain predictions by modifying parts of the input and observing changes in the model's output. SHAP is grounded in game theory and provides consistent feature importance estimates, but its computational cost becomes high for image-based data. This can limit its practical use in real-time or large-scale medical imaging applications.

LIME offers a more practical and intuitive alternative. It generates local explanations by approximating the complex model's behaviour for a single input using a simple, interpretable model. In the case of images, LIME operates on superpixels, which correspond to coherent image regions and are easier for humans to interpret. This allows LIME to clearly highlight areas that positively or negatively influence the model’s decision.

The choice of LIME in this work was motivated by several factors. First, LIME is fully model-agnostic and can be directly applied to YOLOv8 without modifying the detection architecture. Second, its focus on local explanations aligns well with clinical requirements, where understanding individual predictions is more important than global model behaviour. Finally, LIME provides both visual explanations and numerical importance values, enabling not only qualitative assessment but also quantitative evaluation of explanation stability and relevance. For these reasons, LIME was considered a suitable explainability method for interpreting brain tumour detection results in MRI images.

LIME-based explanation of detection results [20]

To enhance the interpretability of the proposed brain tumour detection framework, the Local Interpretable Model-Agnostic Explanations (LIME) method was applied to the YOLOv8 model predictions. LIME is a perturbation-based explainable artificial intelligence technique that provides local explanations for individual predictions by approximating the behaviour of a complex model in the neighbourhood of a specific input instance.

In the context of MRI image analysis, LIME first segments the input image into superpixels, which represent homogeneous, spatially coherent regions. These superpixels serve as interpretable components that can be selectively modified. LIME then generates a large number of perturbed samples by randomly masking different combinations of superpixels, while keeping the remaining regions unchanged. Each

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

perturbed image is subsequently processed by the trained YOLOv8 detection model.

Based on the variations in the model's predictions across these perturbed samples, LIME estimates the contribution of each superpixel to the final detection outcome. A locally weighted linear model is fitted to approximate the complex decision function of YOLOv8 in the vicinity of the analysed prediction. The resulting coefficients of this surrogate model reflect the relative importance of individual superpixels, with positive weights indicating regions that support tumour detection and negative weights indicating regions that reduce the confidence of the prediction.

In this study, LIME was applied specifically to bounding box predictions for detected tumour regions. The generated explanations were visualised by overlaying the most influential superpixels onto the original MRI images, enabling intuitive qualitative assessment of the model's attention. This visualisation enables verification that the model relies on clinically meaningful tumour regions rather than irrelevant background structures.

Evaluation metrics

The performance of the proposed brain tumour detection framework was evaluated using standard object detection metrics and dedicated explainability metrics to assess the quality of LIME-based explanations.

Detection performance was measured using precision, recall, and mean average precision (mAP). Precision is the proportion of correctly detected tumour regions among all detections, while recall measures the model's ability to identify all relevant tumour regions. The mean average precision was computed at an intersection-over-union threshold of 0.50 (mAP@50) as well as across multiple thresholds ranging from 0.50 to 0.95 (mAP@50–95), providing a comprehensive assessment of detection accuracy [21].

To quantitatively evaluate the interpretability of LIME-generated explanations, several complementary metrics were used, each capturing a different aspect of explanation quality. Together, these metrics provide a more comprehensive understanding of how the model distributes attention across image regions and how reliable the generated explanations are.

Let x denote an input MRI image. Using LIME, the image is segmented into N superpixels

$$S = \{s_1, s_2, \dots, s_N\}.$$

Let d denote a single detection produced by YOLOv8, defined by a bounding box and its predicted tumour class.

For explanation purposes, LIME represents perturbed samples using an interpretable binary vector $z \in \{0,1\}^N$, where $z_i = 1$ indicates that the superpixel s_i is present (not masked) and $z_i = 0$ indicates that it is masked.

LIME approximates the behaviour of the detection model in the neighbourhood of image x with respect to detection d using a local linear surrogate model:

$$g(z) = \beta_0 + \sum_{i=1}^{N} \beta_i z_i,$$

where $\beta_i \in \mathbb{R}$ represents the importance weight assigned to superpixel s_i.

Let $\beta = (\beta_1, \beta_2, \dots, \beta_N)$ denote the vector of superpixel importance weights for a single explanation.

Stability measures the robustness of explanations to small perturbations in the input image. For a given image x, LIME explanations are generated multiple times under slight input variations (e.g., noise injection, repeated perturbation sampling).

Let $\beta^{(k)} = (\beta_1^{(k)}, \beta_2^{(k)}, \dots, \beta_N^{(k)})$ denote the vector of superpixel importance weights obtained in the k-th run, $k = 1, \dots, K$.

Stability is defined as the mean standard deviation of superpixel weights across runs:

$$\text{Stability} = \frac{1}{N}\sum_{i=1}^{N} \text{Std}\left(\beta_i^{(1)}, \beta_i^{(2)}, \dots, \beta_i^{(K)}\right).$$

Sparsity quantifies the proportion of superpixels that have a non-negligible contribution to the explanation. Sparse explanations involve only a limited number of influential regions, making them easier to interpret and more clinically meaningful. In contrast, lower sparsity values indicate that the model relies on a broader spatial context.

Let $\tau > 0$ denote a small threshold defining "non-negligible" importance (e.g., $\tau = 0.01 \cdot \max_i |\beta_i|$).

Define the indicator:

$$\mathbb{I}_i = \begin{cases} 1, & |\beta_i| > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Sparsity is computed as:

$$\text{Sparsity} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_i.$$

MaxWeight represents the maximum absolute importance value assigned to a single superpixel.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

This metric highlights the most influential image region contributing to the prediction and provides insight into whether the model relies strongly on a dominant area. A higher MaxWeight value suggests that a specific region plays a critical role in the detection decision, which is desirable when this region corresponds to the actual tumour location.

MaxWeight captures the strength of the most influential superpixel in the explanation:

$$\text{MaxWeight} = \max_{i=1,\dots,N} |\beta_i|.$$

Optionally, normalized by the sum of absolute weights:

$$\text{MaxWeight}_{\text{norm}} = \frac{\max_i |\beta_i|}{\sum_{j=1}^{N} |\beta_j|}.$$

Entropy evaluates how the importance weights are distributed across all superpixels. Lower entropy values indicate that importance is concentrated in fewer regions, leading to clearer, more interpretable explanations. Higher entropy, on the other hand, reflects a more uniform distribution of importance, suggesting that multiple regions jointly influence the model's decision.

First, normalize absolute weights to obtain a probability distribution:

$$p_i = \frac{|\beta_i|}{\sum_{j=1}^{N} |\beta_j|}, i = 1, \dots, N.$$

Entropy is then computed using Shannon entropy:

$$\text{Entropy} = -\sum_{i=1}^{N} p_i \log_2(p_i),$$

where terms with $p_i = 0$ are defined as contributing zero to the sum.

All explainability metrics were first computed at the level of individual detections and then averaged over correctly detected tumour instances in the test set. The practical computation of these metrics was performed under the following assumptions:

- LIME explanations were generated using the same number of superpixels and perturbation samples across all experiments.
- Stability was computed using repeated LIME runs on the same image.
- Sparsity threshold $\tau$was defined relative to the maximum absolute weight to ensure scale invariance.
- Metrics were averaged over correctly detected tumour instances to obtain representative values.

From a reliability engineering perspective, stability relates to robustness under uncertainty (epistemic variability); sparsity reflects importance concentration and cognitive interpretability; MaxWeight aligns with factor dominance in importance analysis; entropy captures uncertainty dispersion across decision factors.

The combination of detection and explainability metrics enables joint evaluation of model performance and transparency, which is essential for assessing the suitability of deep learning models for clinical decision-support applications [13, 15].

Experimental Results

The experimental evaluation was conducted to assess both the detection performance and the explainability of the proposed framework for brain tumour detection from MRI images. The experiments focused on comparing two YOLOv8 model variants and analysing the interpretability of their predictions using the LIME method. All experiments were conducted under identical settings to ensure reproducibility.

Software implementation and Python libraries

The proposed framework was implemented in Python due to its extensive ecosystem of libraries for deep learning, image processing, and explainable artificial intelligence, as well as its widespread adoption in medical image analysis research.

The YOLOv8 detection models were implemented using the Ultralytics library, which provides an efficient and user-friendly framework for training, evaluating, and deploying state-of-the-art object detection models. This library was used for model configuration, training, inference, and evaluation, including the computation of detection performance metrics.

The LIME library was employed to generate explainable visualizations of model predictions. Specifically, the lime_image module was used to produce local explanations based on superpixel perturbations, enabling interpretation of YOLOv8 detection results without requiring access to the internal architecture of the model.

For image preprocessing and superpixel segmentation, the scikit-image library was utilized, particularly the skimage.segmentation module. This library was used to divide MRI images into homogeneous regions required for LIME-based explanations. Additional numerical computations were performed using NumPy, while visualization of results was carried out with Matplotlib, which was used to generate figures, heatmaps, and annotated MRI images.

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

Together, these Python libraries enabled efficient implementation of the proposed detection and explainability pipeline, ensuring reproducibility of experiments and facilitating comprehensive analysis of both model performance and interpretability.

Training setup

The YOLOv8n and YOLOv8s models were trained using the Ultralytics framework on annotated MRI images under identical training conditions to ensure fair comparison. All input images were resized to a fixed resolution of 640 × 640 pixels prior to training.

Both models were trained for 40 epochs with a batch size of 16, using stochastic gradient-based optimization. Standard data augmentation techniques, including geometric transformations and intensity variations, were applied to improve model generalization. The dataset was divided into training, validation, and testing subsets.

Early stopping was applied with a patience of 5 epochs, meaning that training was terminated if no improvement in validation performance was observed over five consecutive epochs. Model weights were automatically saved during training, and the version achieving the best validation performance was selected for final evaluation on the test set.

Detection results

The detection performance was evaluated using two variants of the YOLOv8 object detection model, namely YOLOv8n (nano) and YOLOv8s (small), which differ in terms of model complexity, number of parameters, and computational requirements. The YOLOv8n model represents a lightweight architecture optimized for faster inference and lower hardware demands, whereas YOLOv8s offers increased representational capacity, enabling more accurate feature extraction at the cost of higher computational load.

Both models were trained under identical conditions using the same annotated MRI dataset and the same training and validation splits. To ensure a fair and objective comparison, performance evaluation was conducted on an independent held-out test set that was not used during training or validation. This approach minimises the risk of overfitting and allows assessment of the models' generalisation ability.

Table 1 summarizes the quantitative detection results obtained by the evaluated models. The YOLOv8s model achieved higher precision and recall values compared to YOLOv8n, resulting in improved mean average precision across both evaluated IoU thresholds. This performance gain reflects the ability of the larger model to capture more complex visual patterns present in MRI images.

Table 1. *Detection performance comparison of YOLOv8 models*

Metric	*YOLOv8n*	*YOLOv8s*	*Difference*
			Overall performance
Precision (P)	0.889	0.906	+0.017
Recall (R)	0.863	0.871	+0.008
mAP@50	0.918	0.927	+0.009
mAP@50–95	0.700	0.713	+0.013
			Glioma
Precision (P)	0.802	0.823	+0.021
Recall (R)	0.713	0.768	+0.055
mAP@50	0.812	0.834	+0.022
mAP@50–95	0.519	0.545	+0.026
			Meningioma
Precision (P)	0.945	0.972	+0.027
Recall (R)	0.937	0.937	0.000
mAP@50	0.975	0.983	+0.008
mAP@50–95	0.823	0.836	+0.013
			Pituitary tumour
Precision (P)	0.921	0.922	+0.001
Recall (R)	0.941	0.908	−0.033
mAP@50	0.966	0.962	−0.004
mAP@50–95	0.757	0.758	+0.001

Despite its lower complexity, YOLOv8n demonstrated competitive detection performance, indicating that lightweight models can still provide reliable tumour localization when computational resources are limited. This trade-off between accuracy and efficiency is particularly relevant for practical deployment scenarios, such as real-time clinical decision-support systems.

The confusion matrix presented in Fig. **1** provides a detailed insight into the class-wise performance of the YOLOv8s model for brain tumour detection. Overall, the model demonstrates strong discriminatory capability across all three tumour classes, with many samples correctly classified.

For the glioma class, 282 samples were correctly classified, with only 3 misclassifications as meningioma and no confusion with pituitary tumours. The meningioma class achieved 136 correct predictions, with a small number of errors, including 4 samples misclassified as glioma and 2 as pituitary tumours, indicating partial visual similarity among these tumour types.

The pituitary tumour class showed the most reliable performance, with 184 correctly classified samples and only one misclassification as meningioma. Overall, misclassifications were rare and occurred mainly between

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

glioma and meningioma classes, reflecting overlapping MRI characteristics. These results confirm the robustness and stability of the proposed detection framework.

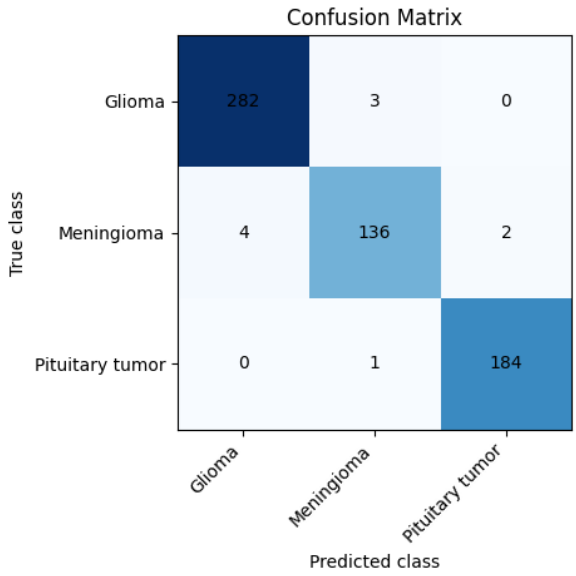

Fig. 1. Confusion matrix of the model

Explainability results

The explainability of the proposed brain tumour detection framework was analysed using the LIME method to identify which image regions contributed most to YOLOv8 model predictions. The objective of this analysis was to verify whether the detection decisions were based on clinically relevant tumour regions and to compare the interpretability of the model variants.

Figure 2 presents an MRI brain image with superpixels highlighted by their contribution to the YOLOv8 model's prediction. Although the tumour region is clearly included among the highlighted areas, LIME also marks several superpixels that do not visually correspond to pathological tissue.

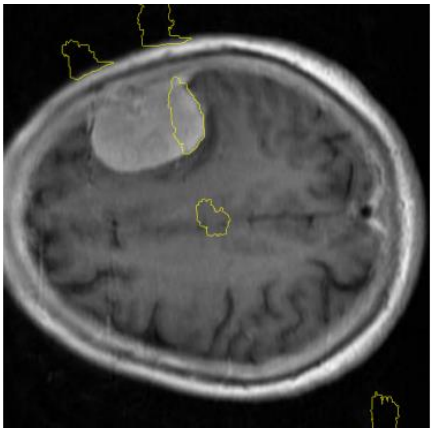

Fig. 2. Visualization of LIME explanation no. 1

This result indicates that the explanation encompasses multiple regions of the image, not all of which represent tumour-related structures, demonstrating the local nature of the LIME-based explanation.

Figure 3 shows the corresponding heatmap visualization of the same LIME explanation. The heatmap provides a continuous representation of feature importance, revealing that the influence of relevant regions is more dispersed compared to well-performing cases. This representation helps to better understand why the explanation is considered less reliable.

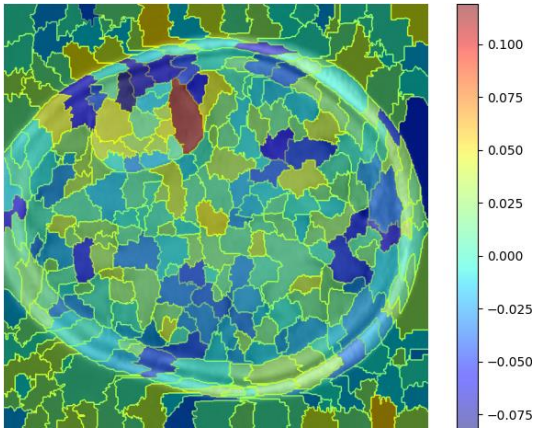

Fig. 3. Heatmap visualization of LIME explanation no. 1

Figure 4 presents an example of a LIME-based superpixel explanation generated for a correctly detected tumour. The highlighted superpixels are predominantly concentrated within the tumour area, indicating that the model focuses on pathological regions when making predictions. The clear alignment between the superpixel importance and the tumour boundaries suggests that the detection model relies on meaningful visual features rather than unrelated background structures.

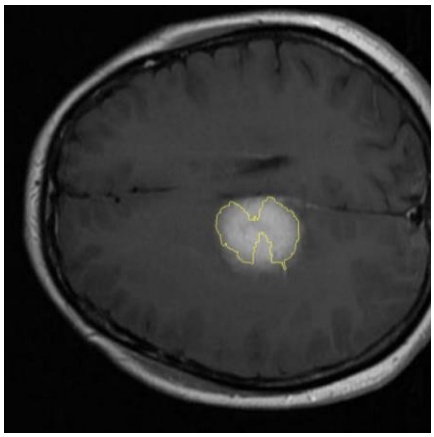

Fig. 4. Visualization of LIME explanation no. 2

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

To provide a more detailed view of the spatial distribution of feature importance, Fig. **5** illustrates a heatmap representation of the LIME explanation for the same detection case. Warmer colors indicate regions with higher influence on the model's prediction. As shown in the figure, the highest contribution values are localised within the tumour region, while the surrounding brain tissue exhibits significantly lower influence. Minor contributions outside the tumour area can be attributed to contextual information or intensity gradients present in MRI images.

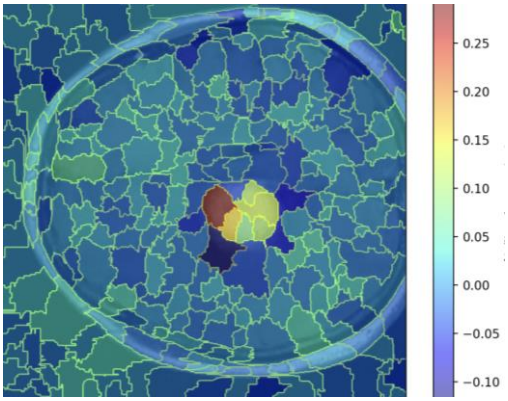

Fig. 5. Heatmap visualization of LIME explanation no. 2

Together, the superpixel-based visualisation (Fig. 4) and the heatmap representation (Figure 5) offer complementary insights into the YOLOv8 model's decision-making process. While the superpixel visualisation highlights discrete influential regions, the heatmap provides a continuous view of importance distribution. This combined analysis confirms that the model's predictions are primarily driven by tumour-related image features, supporting the transparency and reliability of the proposed explainable detection framework.

Qualitative evaluation of LIME explanations

To complement the qualitative analysis based on visual inspections of LIME explanations, a quantitative evaluation was conducted to objectively assess the quality and characteristics of the generated explanations. Quantitative interpretability metrics enable systematic comparison of explanation stability, spatial concentration, and importance distribution, providing deeper insight into the model's decision-making behaviour beyond visual interpretation alone.

Table **2** presents a quantitative evaluation of the explanations generated using the LIME method, providing insight into the stability, concentration, and distribution of superpixel importance. These metrics enable objective assessment of explanation quality beyond qualitative visual inspection.

Table 2. *Quantitative interpretability metrics for LIME*

Metric	***Value***
Stability	0.0302
Sparsity	0.6442
MaxWeight	0.2840
Entropy	5.1498

The stability value of 0.0302 indicates low variability in superpixel weights, suggesting that the generated explanations are robust and consistent under small perturbations of the input image. This behavior is desirable in medical imaging applications, where reliable and repeatable explanations are essential for building trust in automated systems.

The sparsity metric reached a value of 0.6442, indicating that approximately two-thirds of the superpixels contributed meaningfully to the model's decision. This result suggests that the model does not rely on a single isolated region but instead integrates information from a broader spatial context. While highly sparse explanations may indicate overly localized reasoning, the observed sparsity reflects a balanced attention pattern that captures both focal tumour regions and their immediate surroundings.

The maximum superpixel weight of 0.2840 confirms the presence of dominant regions with strong influence on the model output. These highly weighted superpixels correspond to visually salient tumour areas observed in the LIME visualizations, reinforcing the consistency between quantitative metrics and qualitative explanation maps.

Finally, the entropy value of 5.1498 indicates a relatively distributed allocation of importance across multiple superpixels. This suggests that the model's predictions are supported by several relevant image features rather than being dominated by a single region. Such behavior is particularly important in medical image analysis, where pathological patterns may span multiple spatial structures.

Overall, the quantitative results summarized in Table 2 demonstrate that the LIME explanations are stable, informative, and well-aligned with clinically relevant image regions. The combination of focused dominant areas and distributed contextual information supports the transparency and reliability of the proposed explainable brain tumour detection framework.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

Discussion

The results confirm that YOLOv8-based object detection is a suitable approach for brain tumour localization in MRI images. Both evaluated models achieved high detection performance, with YOLOv8s consistently outperforming YOLOv8n across most metrics. The improved performance of YOLOv8s is mainly attributed to its higher model capacity, which enables more effective representation of complex tumour patterns, particularly for gliomas.

Analysis of the confusion matrix showed that misclassifications were infrequent and occurred primarily between glioma and meningioma classes, which exhibit similar visual characteristics in MRI scans. Pituitary tumours were detected with high reliability, likely due to their distinct anatomical location. These findings are consistent with observations reported in related medical imaging studies.

The explainability analysis demonstrated that LIME provides meaningful insight into the model's decision-making process. Both qualitative visualizations and quantitative metrics indicated that the models predominantly focus on clinically relevant tumour regions. The YOLOv8s model produced more stable and concentrated explanations, while YOLOv8n showed more distributed importance patterns, suggesting that increased model complexity contributes not only to higher accuracy but also to improved interpretability.

Several limitations should be considered. The analysis was performed on two-dimensional MRI scans, and tumour localization was limited to bounding boxes rather than precise segmentation. Additionally, LIME provides local explanations that may vary depending on perturbation settings. Despite these limitations, the proposed framework effectively combines detection performance with explainability and supports the development of transparent AI-based medical image analysis systems.

In terms of future directions, the proposed framework can be extended to include a wider range of brain tumour types, which would improve its applicability in more diverse clinical scenarios. Additionally, future work may focus on generating structured textual descriptions of detected or non-detected tumours, providing clinicians with interpretable summaries that complement visual explanations. Such extensions could further enhance the practical usability of explainable detection systems in clinical decision-support workflows.

Conclusions

This paper presented an explainable framework for brain tumour detection from MRI images based on YOLOv8 object detection models and the LIME explainable artificial intelligence method. Experimental results demonstrated that YOLOv8 models achieve reliable detection performance across multiple tumour types, with the YOLOv8s variant providing improved accuracy and more stable explanations.

The integration of LIME enabled transparent interpretation of model predictions by highlighting clinically relevant image regions. Quantitative interpretability metrics further confirmed the stability and informativeness of the generated explanations, supporting the trustworthiness of the proposed approach.

Despite certain limitations, including the use of two-dimensional MRI images and bounding box localization, the proposed framework effectively balances detection accuracy and explainability. Future work will focus on extending the approach to three-dimensional data, incorporating segmentation-based methods, and exploring additional explainability techniques to further enhance clinical applicability.

Acknowledgment

This work was supported in part by the Slovak Research and Development Agency under the grant "Development of a new approach for reliability analysis and risk assessment based on artificial intelligence" (reg.no. APVV-23-0033), and in part by the Ministry of Education, Science, Research, and Sport of the Slovak Republic "Creation of new methods and algorithms of eXplainable Artificial Intelligence based on Importance Analysis" under Grant VEGA 1/0090/25.

Conflict of interest

The authors declare that they have no conflicts of interest, including financial, personal, authorial, or any other nature, that could influence the research or the results published in this article.

Funding

The study was conducted without financial support.

Data availability

The manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technology in the creation of this paper.

References

1. Di Fazio, N., Zanza, C., Longhitano, Y. *et al.* (2026), "Artificial intelligence for early diagnosis in emergency department", *J Anesth Analg Crit Care*, Vol. 6, No.7, 2026, DOI: https://doi.org/10.1186/s44158-025-00334-y
2. Zaitseva, E., Levashenko, V., Rabcan, J., Kvassay, M. (2023), "A New Fuzzy-Based Classification Method for Use in Smart/Precision Medicine", *Bioengineering*, Vol. 10(7), No. 838, DOI: https://doi.org/10.3390/bioengineering10070838
3. Francisco, K.K.Y., Apuhin, A.E.C., *et al.* (2026), "Personalized medicine and health equity: overcoming cost barriers and ethical challenges", *Int J Equity Health* , Vol.25, No. 4. DOI: https://doi.org/10.1186/s12939-025-02710-0
4. Zaitseva, E., Levashenko, V., Rabcan, J., Krsak, E. (2020), "Application of the structure function in the evaluation of the human factor in healthcare", *Symmetry*, Vol.12(1), No. 93. DOI: https://doi.org/10.3390/SYM12010093
5. Zaitseva, E., Levashenko, V. (2026), "Reliability engineering in healthcare: Opportunities and challenges", *Reliability Engineering and System Safety*, Vol. 267, No. 111933. DOI: https://doi.org/10.1016/j.ress.2025.111933
6. Sharma, R.M. (2025), "Artificial intelligence in medical image analysis and molecular diagnostics: recent advances and applications", *J Med Artif Intell.*, Vol.8, 53 p. DOI: https://doi.org/10.21037/jmai-24-412
7. Menze, B.H., *et al.* (2015), "The Multimodal Brain Tumour Image Segmentation Benchmark (BRATS) ", *IEEE Trans. Med. Imaging*, Vol. 34, No. 10, Oct. 2015, pp. 1993–2024. DOI: https://doi.org/10.1109/TMI.2014.2377694
8. Aleid, A., Alhussaini, K., Alanazi, R., Altwaimi, M., Altwijri, O., Saad, A.S., (2023), "Artificial Intelligence Approach for Early Detection of Brain Tumours Using MRI Images", *Applied Science*, Vol. 13, No. 3808. DOI: https://doi.org/10.3390/app13063808
9. Gökmen, N. (2025), "AI techniques for brain tumour segmentation in MRI: a review (2019–2024)", *Netw Model Anal Health Inform Bioinforma*, Vol.14, 168 p. DOI: https://doi.org/10.1007/s13721-025-00650-x
10. Kolarik, M., Sarnovsky, M., Paralic, J., Babic, F. (2023), "Explainability of deep learning models in medical video analysis: a survey", *Peer J Comput. Sci*., Vol. 9, 1253 p, DOI: https://doi.org/10.7717/peerj-cs.1253
11. Kondratenko, Y., Sidenko, I., Kondratenko, G., Petrovych, V., Taranov, M., Sova, I. (2021), "Artificial Neural Networks for Recognition of Brain Tumours on MRI Images". *Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2020.* Communications in Computer and Information Science, Vol 1308. Springer, Cham, DOI: https://doi.org/10.1007/978-3-030-77592-6_6
12. Ali, S., Abuhmed, T., El-Sappagh, Sh., et al. (2023), "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence", *Information Fusion*, Vol. 99, 101805 p. DOI: https://doi.org/10.1016/j.inffus.2023.101805
13. Hassan, S.U., Abdulkadir, S.J., Zahid, M.S.M., Al-Selwi, S.M. (2025), "Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review", *Comput. Biol. Med.*, Vol. 185, 109569 p. DOI: https://doi.org/10.1016/j.compbiomed.2024.109569
14. Zaitseva, E., Levashenko, V. (2025), "Reliability Analysis Based on Aleatory and Epistemic Uncertainty Using Binary Decision Diagrams", *International Journal of Intelligent Systems*, Vol.2025, No. 6471577. DOI: https://doi.org/0.1155/int/6471577
15. Zhou, Z., Hooker, G., Wang, F. (2021), "S-LIME: Stabilized-LIME for Model Explanation", in *Proc. of the 27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining*, pp. 2429-2438. DOI: https://doi.org/10.1145/3447548.3467274
16. Zaitseva, E., Rabcan, J., Levashenko, V., Kvassay, M. (2023), "Importance analysis of decision-making factors based on fuzzy decision trees", *Applied Soft Computing*, Vol. 134, No. 109988. DOI: https://doi.org/10.1016/j.asoc.2023.109988
17. "Medical Image DataSet: Brain Tumour Detection". available: https://www.kaggle.com/datasets/pkdarabi/medical-image-dataset-brain-tumour-detection (last accessed Feb. 04, 2026).
18. Alsufyani, A. (2025), "Performance comparison of deep learning models for MRI-based brain tumour detection", *AIMS Bioeng.*, Vol. 12, No. 1, pp. 1-21. DOI: https://doi.org/10.3934/bioeng.2025001
19. Ultralytics, 'Explore Ultralytics YOLOv8'. available: https://docs.ultralytics.com/models/yolov8/ (last accessed Feb. 04, 2026).
20. Explainable AI (XAI) Using LIME', GeeksforGeeks. available: https://www.geeksforgeeks.org/artificial-intelligence/introduction-to-explainable-aixai-using-lime (last accessed Feb. 04, 2026).
21. Ultralytics, 'Performance Metrics Deep Dive'. available: https://docs.ultralytics.com/guides/yolo-performance-metrics (last accessed Feb. 04, 2026).

Received (Надійшла) 05.01.2026
Accepted for publication (Прийнята до друку) 10.02.2026
Publication date (Дата публікації) 30.03.2026

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

Відомості про авторів / About the Authors

Досталова Крістіна – аспірантка кафедри інформатики, факультет управлінських наук та інформатики, Жилінський університет; Жиліна, Словаччина;
Kristina Dostalova – postgraduate student at the Department of Informatics, Faculty of Management Science and Informatics, University of Žilina; Žilina, Slovakia;
e-mail: kristína.dostalova@gmail.com
ORCID ID: https://orcid.org/0009-0002-8567-716X

Чижмарова Олександра – аспірантка кафедри інформатики, факультет управлінських наук та інформатики, Жилінський університет; Жиліна, Словаччина;
Alexandra Cizmarova – postgraduate student at the Department of Informatics, Faculty of Management Science and Informatics, University of Žilina; Žilina, Slovakia;
e-mail: saskaciz@gmail.com
ORCID ID: https://orcid.org/0009-0007-1021-4722

Клімо Марек – аспірант кафедри інформатики, факультет управлінських наук та інформатики, Жилінський університет; Жиліна, Словаччина;
Marek Klimo – postgraduate student at the Department of Informatics, Faculty of Management Science and Informatics, University of Žilina; Žilina, Slovakia;
e-mail: marek.klimo@fri.uniza.sk
ORCID ID: https://orcid.org/0009-0004-2366-4325
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57226717649

Ярошевич Роман Олександрович – доктор філософії, старший викладач кафедри електронних обчислювальних машин, факультет комп'ютерної інженерії та інформаційних технологій, Харківський національний університет радіоелектроніки; Харків, Україна;
Roman Yaroshevych – PhD, senior lecturer at the Department of Electronic Computers, Faculty of Computer Engineering and Information Technology, Kharkiv National University of Radio Electronics; Kharkiv, Ukraine;
e-mail: roman.yaroshevych@nure.ua
ORCID ID: https://orcid.org/0000-0002-7949-1513
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=58624172500

ВИЯВЛЕННЯ ПУХЛИН ГОЛОВНОГО МОЗКУ НА МРТ-ЗОБРАЖЕННЯХ З МОЖЛИВІСТЮ ПОЯСНЕННЯ ВИКОРИСТАННЯ YOLOV8 У LIME

Актуальність. Точна ідентифікація пухлин головного мозку за допомогою магнітно-резонансної томографії (МРТ) є критично важливим завданням в аналізі медичних зображень. Хоча підходи глибокого навчання часто досягають відмінної продуктивності виявлення, їм часто бракує прозорості в процесах прийняття рішень. Ця відсутність інтерпретованості обмежує довіру в клінічній практиці та створює сильну потребу у впровадженні методів пояснювального штучного інтелекту (XAI). **Об'єкт дослідження**. Об'єктом цього дослідження є автоматизований процес виявлення пухлин головного мозку на МРТ-сканах з використанням підходів на основі згорткових нейронних мереж. **Предмет дослідження**. Предметом дослідження є застосування моделей виявлення об'єктів YOLOv8 разом з методом пояснювальності LIME для інтерпретації виходів моделі в контексті аналізу медичних зображень. **Мета**. Метою цієї статті є розробка та оцінка пояснювальної структури для виявлення пухлин головного мозку, яка інтегрує виявлення об'єктів на основі YOLOv8 з інтерпретованістю, керованою LIME, за підтримки кількісної оцінки якості пояснення. **Результати**. Експериментальна оцінка демонструє, що моделі YOLOv8 здатні точно виявляти вибрані типи пухлин головного мозку на зображеннях МРТ, тоді як LIME успішно визначає області зображення, які мають найбільший вплив на рішення моделі. Запропоновані кількісні метрики підтверджують стабільність та розрідженість згенерованих пояснень, тим самим покращуючи інтерпретованість та надійність запропонованої системи виявлення.

Ключові слова: виявлення пухлини головного мозку; МРТ; YOLOv8; зрозумілий штучний інтелект; LIME; аналіз медичних зображень.

Бібліографічні описи / Bibliographic descriptions

Досталова, К., Чижмарова, О., Клімо, М., Ярошевич Р.О. Виявлення пухлин головного мозку на МРТ-зображеннях з можливістю пояснення використання YOLOv8 у LIME. *Сучасний стан наукових досліджень та технологій в промисловості*. 2026. № 1 (35). С. 17–27. DOI: https://doi.org/10.30837/2522-9818.2026.1.017

Dostalova, K., Cizmarova, A., Klimo, M., Yaroshevych, R. (2026), "Explanation Detected Brain Tumours in MRI Images Using YOLOv8 with LIME-Based Interpretation", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (35), P. 17–27. DOI: https://doi.org/10.30837/2522-9818.2026.1.017