Automated Classification and Segmentation of Brain Tumors on MRI Images Using Superpixel-Based Machine Learning

Nadezhda Shchegoleva[1](✉), Nadezhda Pronina[1], Nataliya Zalutskaya[2], and Jasur Kiyamov[3]

[1] Saint Petersburg State University, St.Petersburg, Russia
n.shchegoleva@spbu.ru, st105895@student.spbu.ru
[2] Federal State Budgetary Institution «Bekhterev National Medical Research Psychiatry and Neurology Center», Ministry of Health of the Russian Federation, St. Petersburg, Russia
[3] Samarkand Branch of Tashkent University of Information Technologies, Tashkent, Uzbekistan
z.kiyamov@spbu.ru

Abstract. Brain tumor detection remains one of the most pressing challenges in medical image analysis due to the complexity and variability of tumor structures. This study presents an automated approach for analyzing magnetic resonance imaging (MRI) scans, aimed at identifying pathological cases and localizing abnormal regions. A hybrid classification model based on support vector machines and k-nearest neighbors is used to distinguish between normal and pathological images. For segmentation, a superpixel-based method is applied to highlight tumor areas. The system combines traditional image preprocessing with statistical and textural feature extraction to enhance diagnostic accuracy. Experimental results confirm the effectiveness of the proposed two-stage pipeline in supporting early diagnosis and reducing the cognitive workload of clinicians.

Keywords: brain tumor · MRI · hybrid classifier

1 Introduction

Brain tumors represent one of the most complex and dangerous pathologies of the central nervous system. According to the World Health Organization (WHO), 20 million new cancer cases were registered worldwide in 2022, with 9.7 million cancer-related deaths. Brain tumors remain among the most challenging forms of cancer, requiring early diagnosis and treatment (WHO).

According to recent statistical data published on the "Main Oncological Portal of Russia" [8], the share of malignant brain tumors in the structure of oncological diseases among the male population in 2023 was approximately 1.4%, and 1.2% among women. Although these figures are lower compared to other cancer localizations (such as prostate cancer in men or breast cancer in women), the dynamics from 2013 to 2023 indicate an increase in the incidence of central nervous system (CNS) tumors. In particular, the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15894, pp. 264–281, 2026.
https://doi.org/10.1007/978-3-031-97648-3_18

estimated average annual growth rate for this group of diseases reaches approximately 0.9–1%.

Despite advances in modern medicine, the diagnosis of brain tumors remains a challenging task. Magnetic resonance imaging (MRI) is the gold standard for brain imaging, but interpreting the results requires highly qualified specialists and significant time investment. Furthermore, human factors, physician fatigue, and image quality can negatively affect diagnostic accuracy.

In this context, the development of automated methods for analyzing medical images, utilizing machine learning and computer vision technologies, becomes especially relevant. These algorithms make it possible to identify pathological changes at early stages, minimize the influence of subjective factors, and increase the speed and accuracy of diagnosis.

Analyzing brain tumor structures is a complex task, as not all formations have clear boundaries and distinct visual features. In such cases, traditional diagnostic methods may not be sufficiently accurate, which increases the importance of algorithms that analyze the textural and morphological features of images.

2 Overview

2.1 Research Objectives

Main Objective. The primary goal of this study is to develop an automated system that, in the first stage, determines the presence or absence of pathological changes in the brain, and in the second stage, if such changes are detected, identifies their localization. This two-stage approach to MRI result processing helps optimize the required computational resources and, in some cases, simplifies the interpretation of the results.

The research tasks aimed at achieving this goal can be formulated as follows:

1. Data Preparation. Collection and analysis of an MRI image dataset in compliance with licensing requirements. Preliminary image preprocessing: noise removal, contrast and brightness adjustment, histogram equalization.
2. Classification. Analysis of existing machine learning methods and identification of their advantages and disadvantages for solving the tumor detection task (binary classification: "normal" or "pathology"). Comparison of various algorithms and evaluation of result quality using accuracy, recall, and F1-score metrics.
3. Segmentation of Detected Pathologies. Analysis of existing segmentation algorithms and evaluation of their suitability for detecting pathologies (thresholding methods, superpixels). Generation of masks for pathological areas and assessment of segmentation quality using appropriate metrics.
4. Comparative Analysis and Evaluation of Results. Comparison of results based on selected metrics and search for the optimal combination of "classifier → segmentation method".

2.2 Literature Review

The automation of detection and segmentation of low-grade gliomas in brain MRI scans remains one of the most in-demand tasks in the field of medical imaging. Various approaches to solving this problem are reflected in scientific studies, where each

group of authors proposes their own set of algorithms and methods, tailored to different aspects of image processing, computational capabilities, and the specific characteristics of medical data.

One of the earliest fundamental approaches to image segmentation is the work by Felzenszwalb and Huttenlocher [4], in which the authors described a graph-based model that allows for clustering pixels based on proximity measures and controlling the "breaks" within potential segments. The proposed method is characterized by relatively high processing speed and flexible tuning of the scale parameter, which determines the level of detail in the final segmentation. In the context of brain MRI analysis, this algorithm can be applied for preliminary separation of homogeneous regions; however, its effectiveness highly depends on proper hyperparameter selection and preliminary image preprocessing.

The work of Kalavathi [7] is devoted to studying a multi-threshold modification of the Otsu algorithm. The application of this method demonstrated that using multiple thresholds allows for more accurate differentiation of various brain tissues—white matter, gray matter, and cerebrospinal fluid. However, when it comes to heterogeneous tumor regions with low contrast, this approach may lead to inaccurate results, and additional procedures (such as clustering or machine learning) are required to improve accuracy.

At the classification stage—i.e., determining the presence of pathological tissue—hybrid algorithms often prove to be effective solutions. For example, Machhale et al. [5] showed that a combination of support vector machines (SVM) and k-nearest neighbors (KNN) outperforms each of these techniques individually when detecting tumors in MRI scans. The SVM-KNN hybrid takes into account both the global feature structure and their local neighborhood, which is especially important when analyzing highly variable medical data.

At the same time, the emergence of large open image datasets and the increase in computational power have driven the development of neural network-based methods. Menze and colleagues [6] created the benchmark BRATS, a reference dataset for evaluating brain tumor segmentation algorithms across multiple MRI modalities (T1, T1c, T2, FLAIR). This study demonstrated the advantages of deep convolutional neural networks (CNNs), especially when sufficient data and diverse contrast types are available. The authors thoroughly examine the Dice Similarity Coefficient and Intersection over Union (IoU) metrics [6], as well as challenges encountered when processing images of varying quality and resolution.

An important aspect of modern oncological diagnostics is radiogenomics—a field focused on correlating radiological and genetic characteristics of tumors. Buda, Saha, and Mazurowski [3] showed that certain shape parameters derived from deep image analysis models correlate with molecular subtypes of low-grade gliomas. A similar conclusion was reached in the work by Pedano et al. [2], where data from The Cancer Genome Atlas (TCGA) were analyzed alongside MRI results. Both studies confirm that accurate segmentation and detailed extraction of features describing the shape of a tumor can provide additional insights into the tumor's nature, which in turn can influence prognosis and treatment planning.

Thus, the analysis indicates that successful neoplasm diagnosis requires three key components: a reliable classification system (such as SVM-KNN or Random Forest), an

accurate segmentation method tailored to data quality (e.g., Felzenszwalb's algorithm, SLIC, or U-Net), and the integration of imaging features with genetic information to enhance the understanding of tumor characteristics and improve treatment strategies.

2.3 Overview of Classification Approaches

The task of classification in the context of medical image analysis typically involves making a binary decision: "tumor present" or "no tumor." The accuracy of this step largely determines the effectiveness of subsequent processing stages, as it allows either the immediate exclusion of healthy cases or the forwarding of images with signs of pathology for more detailed processing (segmentation). Several main groups of methods are commonly used for classification in medical applications:

Traditional Methods.

Support Vector Machine (SVM). SVMs perform well with high-dimensional data and can construct both linear and non-linear decision boundaries by using appropriate kernels (e.g., RBF, polynomial). The method focuses on examples located near the decision boundary, enabling it to build robust models even with a limited amount of data, especially when classes are well-separated. However, the effectiveness of SVM greatly depends on proper hyperparameter tuning.

k-Nearest Neighbors (KNN). KNN assigns a data point to the class most common among its k nearest neighbors in the feature space. The method is simple to understand and implement but may become relatively slow with large datasets. Additionally, KNN is sensitive to feature scaling and the choice of distance metric.

Ensemble Methods (Random Forest, Gradient Boosting).

Ensemble algorithms build multiple decision trees, each trained on a random subset of the data. In the case of Random Forest, the final result is determined by the majority vote of all trees, which reduces overfitting and increases robustness to noise. A major advantage of this approach is its high interpretability, particularly when analyzing feature importance.

Hybrid Approaches.

SVM-KNN. Hybrid methods combine the strengths of multiple algorithms—for example, SVM and KNN. The data is first separated using the support vector machine, and then information about the nearest neighbors (KNN) is used to refine the boundaries between classes or to handle ambiguous cases near the decision margin. According to several experiments (e.g., Machhale et al. [5]), this approach often yields higher accuracy and F1-scores than using either SVM or KNN alone.

Modern Neural Networks.

In recent years, convolutional neural networks (CNNs) have gained popularity due to their ability to automatically extract the most relevant features, minimizing the need for expert-driven feature selection. Convolutional layers are capable of capturing spatial relationships between pixels in brain images, which is particularly important in the medical domain. However, training neural networks requires large volumes of labeled data and significant computational resources.

Thus, the choice of classification method depends on data volume, processing speed requirements, and computational resources. Traditional methods offer predictability and

transparency, while hybrid and neural network approaches can achieve higher accuracy if properly tuned and supported by sufficient training data.

2.4 Overview of Segmentation Approaches

Once the presence of a tumor has been established, the next step is to determine its exact location and size—that is, segmentation and delineation of the pathological region. The complexity of this task lies in the heterogeneity of MRI scans, varying tissue contrast, and possible noise, artifacts, and variations in brain characteristics across different patients. Below are the main groups of segmentation algorithms commonly used in medical image analysis.

Thresholding Methods.

Otsu's Algorithm (and its modifications). A classic thresholding segmentation technique where the optimal threshold is selected to maximize inter-class variance between pixel intensities. Multi-threshold extensions (e.g., Kalavathi, 2013 [7]) allow for the differentiation of multiple tissue types (e.g., white matter, gray matter, cerebrospinal fluid). However, in the presence of tumors whose contrast may overlap with that of normal tissues, thresholding methods often produce results that require refinement.

Clustering- and Superpixel-Based Methods.

SLIC (Simple Linear Iterative Clustering). SLIC divides the image into compact superpixels—small regions grouping adjacent pixels with similar visual characteristics, such as color or brightness. Superpixels significantly reduce the dimensionality of the task: instead of pixel-level analysis, groups of similar pixels are considered. In medical applications, this approach simplifies classification by aggregating homogeneous regions. However, the final result is highly dependent on proper parameter selection (e.g., the number of superpixels, compactness factor).

Felzenszwalb's Method. A graph-based segmentation algorithm where pixels are treated as vertices connected by similarity-based edges (brightness, color, spatial coordinates). Vertices are merged into clusters if the gap between them exceeds a threshold. The method offers high computational efficiency (O(n log n)) [4] and can detect heterogeneous regions, but its performance is sensitive to the scale parameter and input image characteristics.

Neural Network-Based Algorithms.

U-Net and FCN. Segmentation architectures based on convolutional neural networks are widely used in medical applications (e.g., Menze et al., 2015 [6]). Neural networks can automatically detect complex patterns but require large amounts of labeled data and significant computational resources. With sufficient training data, they often achieve the highest scores on metrics such as Dice and IoU.

In practice, the choice of segmentation approach is usually determined by data quality (e.g., presence of contrast or noise), research goals (e.g., preliminary segmentation vs. precise boundary extraction), and available computational resources. Classical and graph-based methods are often sufficient for cases with clear contrast and limited data, while neural networks offer superior accuracy when large, well-annotated datasets are available.

3 Description of Methods

This chapter is dedicated to a detailed examination of the methods for automatic analysis of brain MRI images, as well as the justification for parameter choices at each processing stage. The system developed in this study is based on a two-stage scheme: classification (determining the presence of a tumor) and segmentation (precise delineation of the pathological region) [5, 7].

The implemented system follows these data processing stages:

Preprocessing (noise removal, gamma correction, morphological operations).

Feature extraction and classification (detection of the presence of a tumor).

Segmentation (refining the tumor boundaries if the classifier result is positive).

Postprocessing of the final mask (morphological operations, noise removal).

3.1 Image Preprocessing

Noise Removal and Gamma Correction. The system receives MRI images in.tif format, which may potentially contain a high level of noise [1]. To improve image quality, the following algorithms are most commonly used:

Median filter (with a 3×3 or 5×5 window), which replaces the value of each pixel with the median of its surrounding neighborhood:

$$I^{I}(x, y) = \text{median}\ \{I(i, j) | (i, j) \epsilon N(x, y)\} \tag{1}$$

where N(x,y) is the neighborhood window around the pixel (x,y), and III is the original image. This approach reduces salt-and-pepper noise without significantly blurring the edges [5].

Gamma correction.

$$I^{I}(x, y) = I(x, y)^{\gamma} \tag{2}$$

where γ is a parameter that determines the degree of brightness enhancement. In this study, γ values in the range of 0.5–2.0 were used to adjust low-contrast regions of the image. Lower values ($\gamma < 1\backslash gamma < 1\gamma < 1$) increase the contrast in darker areas of the image, which facilitates tumor detection against the background of healthy tissue [7].

Morphological Operations and Artifact Removal. After gamma correction, morphological operations such as erosion and dilation are typically applied to remove small noise artifacts or, conversely, to fill in small gaps within the tumor volume. Specifically, the opening and closing methods from the scikit-image library were used. The first is defined as erosion followed by dilation, and the second as dilation followed by erosion [9]. Erosion and dilation are defined by a kernel (structuring element B) and can be described by the following formulas [10]:

Erosion:

$$(A \ominus B) = \{z \,|B_z \subseteq A\} \tag{3}$$

Dilation:

$$(A \oplus B) = \{Z | B_{-Z} \cap A \neq \varnothing\} \tag{4}$$

Here, A is a set (a binary mask), and B_z is a shifted copy of the structuring element.

3.2 Feature Extraction and Classification

Feature Vector Construction. To solve the classification task, it is necessary to extract the most informative features from the images. The main groups of features used in this work are listed below:

1. Statistical characteristics [6]:

The most basic and widely used measure is the **mean**, defined by the formula:

$$\text{mean} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5}$$

This parameter reflects the average brightness level in the analyzed region. However, it can be sensitive to outliers (very bright or very dark pixels).

Variance characterizes the spread of values around the mean:

$$\text{variance} = \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \text{mean})^2 \tag{6}$$

This parameter indicates how much the pixel intensity values deviate from the mean.

Standard Deviation (STD) is the square root of the variance:

$$\text{std} = \sigma = \sqrt{\text{variance}} \tag{7}$$

Higher values of variance and standard deviation indicate that the brightness levels in the analyzed image region are more heterogeneous (variable).

Median

$$\text{median} = \begin{cases} x_{\left(\frac{N+1}{2}\right)}, \text{ if N odd} \\ \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2}, \text{ if N even} \end{cases} \tag{8}$$

where x_k is the k-th element in the sorted set $\{x_1, x_2, ..., x_N\}$. The median represents the central value that divides the sample into two equal parts. It is less sensitive to outliers compared to the mean.

Entropy is estimated based on the image intensity histogram. Let the histogram be divided into M intervals (bins), and p_k be the probability that a pixel falls into the k-th bin (the sum of all p_k is 1). Then entropy is defined as:

$$\text{entropy} = -\sum_{i=1}^{M} p_k \log_2 p_k \tag{9}$$

Entropy measures the level of "uncertainty." A high value indicates that the region contains a wide range of intensity values (i.e., it has a complex structure).

Skewness. This measures the asymmetry of the distribution relative to the mean (left-skewed or right-skewed). A common formula (without bias correction) is:

$$\text{skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean})^3}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean})^2)^{\frac{3}{2}}} \tag{10}$$

If skewness > 0, the distribution has a longer tail on the right, indicating a tendency toward brighter pixels. If skewness < 0, the tail is shifted to the left.

Kurtosis evaluates how "heavy" the tails of the distribution are compared to a normal (Gaussian) distribution. There are several formulas; the most common one (Pearson's definition) gives a value of 0 for a normal distribution.

$$\text{kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean})^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean})^2)^2} - 3 \tag{11}$$

2. Textural features – Haralick features based on the gray-level co-occurrence Matrix (GLCM) [4]. For example:

$$C = \sum_{i=1}^{L-1}\sum_{j=1}^{L-1}(i-j)^2 p(i,j) \tag{12}$$

where $p(i, j)$ is the probability of the co-occurrence of intensity values i and j, and L is the number of gray levels. Other features such as Energy, Homogeneity, Correlation, etc., are computed in a similar way.

All these features are combined into a feature vector x, which is then normalized, for example, using the StandardScaler method:

$$x_{scaled} = \frac{x - \mu_x}{\sigma_x} \tag{13}$$

where σ_x and μ_x are the mean and standard deviation of the feature x across the training dataset.

3. LBP features (Local Binary Patterns) is another widely used method for texture description, computed at the neighborhood level for each pixel [7]. Let P be the number of sampling points around the center, and R the radius of the circle. Then for each pixel at coordinates (x_c, y_c). Determine the intensity of the central point:

$$I_c = I(x_c, y_c) \tag{14}$$

Select P neighboring points I_p, $p \epsilon \{0, ..., P-1\}$, lying within the radius R.
Compare each neighboring intensity I_p with the central intensity I_c:

$$s(I_p - I_c) = \begin{cases} 1, \text{if} I_p \geq I_c \\ 0, \text{else} \end{cases} \tag{15}$$

Form a binary number where each bit corresponds to the result of the comparison $s(I_p - I_c)$ for the corresponding neighboring pixel:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(I_p - I_c)2^p \tag{16}$$

The resulting code (usually $P = 8$ or $P = 16$) is interpreted as the index of a bin in a histogram. Thus, for the entire ROI (Region of Interest) or image, an LBP histogram is constructed, which describes local texture patterns (such as edges, corners, spots). Since the comparison is done bitwise, the method is less sensitive to global changes in lighting or contrast.

Textural features allow for a deeper analysis of the internal structure of the tumor and surrounding tissues, in addition to statistical characteristics (mean, std, etc.). GLCM/Haralick metrics analyze the co-occurrence of intensity values in the pixel neighborhood, making it possible to detect texture changes. LBP reflects micro-local changes in brightness, while Gabor filters are aimed at detecting structural features (lines, contours) in specific orientations. All of these methods together form a sufficient set of features that serves as a basis for more accurate and reliable decisions in brain tumor classification and segmentation tasks [4, 7].

3.3 Classification Algorithms

Support Vector Machine (**SVM**) is one of the classic and reliable machine learning algorithms, well-proven in medical applications [5]. The core idea is to find a hyperplane that best separates objects of one class from those of another, taking into account the margin. The margin is defined as the distance from the separating hyperplane to the closest objects from both classes (support vectors); maximizing this distance improves the model's generalization ability and robustness to noise.

The mathematical formulation in the linear case is:

$$\min \frac{1}{2}||w||^2 \text{ subject to } y_i(w^T x_i + b) \geq 1, i = 1, ..., N. \tag{17}$$

where $x_i \in R^d$ is the feature vector of the i-th object, $y_i \epsilon \{-1; +1\}$ is the class label (typically 0/1 in medicine, though it can be reformatted to $-1/+1$), and w and b define the position of the separating hyperplane in the feature space. During training, values of w and b are selected such that the margin between classes on the training data is maximized.

In the nonlinear case, kernel functions are introduced. The most commonly used is the RBF kernel:

$$K(x_i, y_j) = \exp(-\gamma \left|\left| x_i - y_j \right|\right|^2) \tag{18}$$

which allows the construction of complex decision boundaries between classes. Key hyperparameters of SVM include:

C (regularization): controls the balance between maximizing the margin and minimizing classification errors; γ (for the RBF kernel): defines the "influence radius" of individual training samples.

k-Nearest Neighbors (**KNN**) is an algorithm that classifies a new object based on the majority class among its k nearest neighbors in feature space [5, see Sect. 2.3]:

$$\hat{y} = \text{majority}\{y_{(1)}, y_{(2)}, ..., y_{(k)}\} \tag{19}$$

where $y_{(j)}$ is the class of the j-th nearest neighbor. The main parameter is the number of neighbors k. If k is too small, the algorithm may overfit; if too large, it loses local sensitivity.

Random Forest is an ensemble of decision trees [6]. Each tree is trained on a random subset of samples and features, and the final class label is determined by majority voting:

$$\hat{y} = \text{majority}\{\hat{y}_1, ..., \hat{y}_T\} \tag{20}$$

where T is the number of trees in the forest. Key parameters include: n_estimators – number of trees; max_depth – maximum allowed depth of a tree; max_features – number of features to consider at each split. This algorithm is resistant to overfitting and often demonstrates high accuracy when hyperparameters are properly tuned.

3.4 Segmentation Methods

The **SLIC** (Simple Linear Iterative Clustering) algorithm partitions an image into N superpixels, taking into account not only similarity in brightness l, but also spatial coordinates (x,y). Each cluster is represented by a center $c_k = (l_k, x_k, y_k)$. During the iterative process, cluster centers are updated to minimize the combined distance:

$$D(c_k, p) = \sqrt{\left(\frac{l_k - l_p}{\lambda_l}\right)^2 + \left(\frac{x_k - x_p}{\lambda_s}\right)^2 + \left(\frac{y_k - y_p}{\lambda_s}\right)^2} \tag{21}$$

where λ_l is the intensity (brightness) distance, λ_s is the spatial distance [4]. In medical applications, SLIC is often used to reduce analysis to regional (superpixel-based) rather than pixel-level processing.

To evaluate the quality of segmentation, Dice and Intersection over Union (IoU) metrics are used [6].

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{22}$$

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{23}$$

where A is the ground truth tumor mask, and B is the predicted region by the algorithm [2, 3]. Values above 0.8–0.9 are generally considered to indicate good segmentation performance.

3.5 Parameter Selection and Tuning

Classification. Optimal hyperparameters (e.g., number of trees for Random Forest, values of γ and C for SVM, number of neighbors k for KNN, etc.) were selected using 5-fold

cross-validation and grid search [5, 7]. The metrics Accuracy, Precision, and Recall were computed as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{24}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{25}$$

$$\text{Reccal} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{26}$$

where TP (True Positives) is the number of tumor-containing images correctly classified as "pathology"; TN, FP, and FN denote the number of true negatives, false positives, and false negatives, respectively.

For SLIC, the number of superpixels (n_segments) ranged from 100 to 300, and the compactness parameter from 5 to 30. After generating the binary tumor mask, hole filling (binary_fill_holes) and removal of small connected components (smaller than 50–100 pixels) were applied. Final parameters were selected experimentally on a training subset, and results were evaluated using the Dice coefficient and Intersection over Union (IoU) metrics [2, 6].

3.6 Most Promising Solution

Novelty of the Proposed Approach. The proposed method implements a two-stage pipeline that combines a hybrid SVM-KNN classifier with superpixel-based segmentation (SLIC) to optimize brain tumor detection on MRI scans. Unlike traditional single-step approaches [5, 6], our system first filters healthy cases, reducing computational load, and applies segmentation only to suspected pathological images. By integrating classical statistical (mean, variance, entropy) and textural (Haralick, LBP) features [4, 6] extracted from superpixels, the system enhances robustness and interpretability compared to deep learning models that require large labeled datasets [6]. This approach provides a balance between efficiency and diagnostic accuracy, making it particularly suitable for clinical settings with limited data.

Original Dataset. In this study, the "LGG MRI Segmentation" dataset published on the Kaggle platform by Mateusz Buda [1] was used. The dataset is distributed under the CC BY-NC-SA 4.0 license and contains brain MRI scans for 110 patients from the TCGA-LGG collection stored in The Cancer Imaging Archive (TCIA). The original tomographic data were obtained as part of The Cancer Genome Atlas (TCGA) project and are accompanied by corresponding genetic and clinical information [2]. A full description of the dataset and related studies is available in the author's publications [3]. The Kaggle version of "LGG MRI Segmentation" includes manually labeled masks of pathological areas. To comply with licensing and TCGA requirements, it is recommended to cite the TCGA-LGG collection [2] when referencing this dataset and to follow the citation guidelines outlined on the TCGA Research Network website. An example of an original MRI scan in.tif format is shown in Fig. 1 (e.g., TCGA_CS_5393_19990606_8).

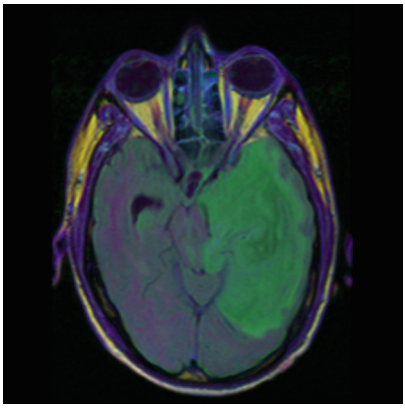


Fig. 1. MRI-scan



Fig. 2. Corresponding tumor annotation mask of the MRI scan

The corresponding manually created tumor annotation mask is shown in Fig. 2 (Example: TCGA_CS_5393_19990606_8_mask).

The analysis of processing methods and the characteristics of experimental data obtained during this study led to the implementation of the two-stage scheme "classification → segmentation" as the optimal approach. The study employed both classical algorithms (Random Forest) and a hybrid model (SVM-KNN). According to the experimental results, the hybrid method demonstrated high accuracy with a limited amount of data, taking into account both the global structure of the feature space (via SVM) and local characteristics (via KNN). At this stage, the SLIC (Simple Linear Iterative Clustering) algorithm was applied, allowing the original image to be reduced to a set of superpixels.

3.7 Results Analysis

Baseline Comparison. To evaluate the effectiveness of the proposed system, simple baseline methods were tested on the same dataset. The visual results reveal key limitations of the Otsu thresholding approach. In the first example (Fig. 3), the algorithm falsely detects large anatomical regions as tumor tissue, despite the absence of any pathology, resulting in a false positive. In the second case (Fig. 4), although a tumor is clearly visible and correctly annotated in the ground truth, Otsu fails to segment it and instead highlights peripheral skull structures—a typical false negative. These outcomes illustrate that global thresholding techniques are insufficiently reliable for medical image segmentation, particularly in cases with low contrast or complex morphology, reinforcing the need for more robust and context-aware hybrid methods.



Fig. 3. Example of misclassification by Otsu: the true mask is empty, but the thresholding method incorrectly detects large anatomical regions as tumor.



Fig. 4. False negative by Otsu: the method misses the tumor and segments only peripheral structures.

Segmentation Results and Analysis. Experiments have shown that the two-stage data processing procedure "classification → segmentation" offers the following advantages:

reduced segmentation time for images that do not contain tumors, due to early filtering of standard (healthy) cases.

Based on extracted statistical (Mean, Std, Entropy, Skewness, etc.) and textural (Haralick, LBP) features, several machine learning methods were trained, including SVM, KNN, Random Forest, and a hybrid SVM-KNN model.

According to the experimental results, the hybrid SVM-KNN model demonstrated high accuracy and balanced predictions, even in the presence of class imbalance in the dataset. The final confusion matrix and classification metrics are shown in Table 1 and Table 2, respectively.

Table 1. Confusion matrix for SVM-KNN.

	Predicted 0	Predicted 1
True 0	511	0
True 1	1	274

Table 2. SVM-KNN classification metrics.

	0	1
Precision	0.998	1
Recall	1	0.996
F1-score	0.999	0.998
Support	511	275

Overall Accuracy: 0.999.
Macro-averaged: Precision = 0.999, Recall = 0.998, F1-score = 0.999.

The model's high sensitivity to pathological cases (Recall = 0.996) indicates that virtually no tumors were missed—only one out of 275 cases was misclassified. At the same time, specificity on healthy images remained perfect: all 511 class 0 examples were correctly classified, meaning no false positives were detected. This behavior is particularly important in the context of medical diagnostics, where the cost of error is extremely high. The results demonstrate that the classifier can be effectively used as a primary filter: healthy images are excluded without requiring further segmentation, while suspicious cases proceed to the next processing stage, optimizing computational resource usage.

As part of the comparative analysis, the Random Forest model was also tested and showed perfect performance on the test set. The algorithm correctly classified all examples, with zero errors in both healthy and pathological cases. The final confusion matrix and classification metrics are shown in Table 3 and Table 4.

The perfect match between predictions and ground truth confirms that the Random Forest model performed excellently on this binary classification task. Its high robustness

Table 3. Confusion matrix for Random Fores

	Predicted 0	Predicted 1
True 0	511	0
True 1	0	275

Table 4. Random Forest classification metrics.

	0	1
Precision	1	1
Recall	1	1
F1-score	1	1
Support	511	275

Overall Accuracy: 1.000
Balanced Accuracy: 1.000
ROC-AUC: 1.000
Class-averaged: Precision = 1.000, Recall = 1.000, F1-score = 1.000

to noise, interpretability, and ability to evaluate feature importance make it an attractive choice—especially on clean, balanced datasets.

However, it should be noted that such ideal results may be due to the quality and structure of the dataset. In real-world scenarios, where data is more diverse and may contain artifacts or borderline cases, model performance may vary. Nonetheless, the obtained results highlight the strong potential of Random Forest as a classifier for medical imaging tasks, particularly during the initial screening and diagnostic stages.

In the segmentation phase, the Random Forest classifier was combined with the SLIC algorithm, and experiments were conducted to evaluate segmentation performance with varying numbers of clusters (2, 3, and 4), while the number of segments was fixed at 380, and compactness at 14.

The segmentation results demonstrated the following performance metrics:

2. Clusters: dice Coefficient (mean): 0.0374; IoU (mean): 0.0198 (Fig. 5, Fig. 6, Fig. 7)

3. Clusters: dice Coefficient (mean): 0.0182; IoU (mean): 0.0099 (Fig. 8)

4. Clusters: dice Coefficient (mean): 0.0125; IoU (mean): 0.0069 (Fig. 9)

These relatively low segmentation metrics indicate that, while the Random Forest classifier perfectly identifies the presence or absence of tumors at the classification stage, the current approach to segmentation using the SLIC algorithm needs further improvement. However, the current segmentation approach consistently includes significant background areas along with the targeted tumor regions. Specifically, the segmentation method does not adequately differentiate between relevant brain structures and irrelevant background or skull areas.

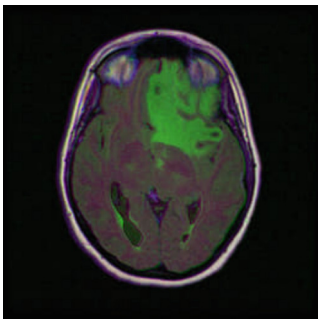


Fig. 5. MRI-scan



Fig. 6. Results using 2 clasters



Fig. 7. Corresponding tumor annotation mask of the MRI scan



Fig. 8. Results using 3 clasters



Fig. 9. Results using 4 clasters

4 Conclusions

The developed automated system significantly accelerates the analysis of medical images, helping doctors promptly detect pathological changes in brain MRI scans. At this stage of research, the system already successfully identifies neoplasms, substantially easing the doctor's workload and enabling faster examination of a greater number of patients. In the future, we plan to improve the algorithm by excluding cranial bones from the analysis process. While the current superpixel-based segmentation demonstrates promising results, further research involving clinical validation and evaluation on larger, more diverse datasets is necessary to fully establish the robustness and diagnostic reliability of this approach.

However, regardless of automation level, the final diagnosis must always be made by a medical professional—relying not only on the results of automatic MRI analysis, but also on the original scans and additional tests such as blood work, EEG, the Wechsler test, and others.

Thus, the proposed approach is already enhancing the efficiency of medical diagnostics today, enabling the timely detection of tumors and the initiation of necessary treatment.

References

1. Buda, M.: LGG MRI Segmentation (Brain MRI Segmentation). Kaggle, License: CC BY-NC-SA 4.0 (2019). https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation/data. Access 28 Mar 2025
2. Pedano, N., Flanders, A.E., Scarpace, L., Mikkelsen, T., Eschbacher, J., Hermes, B., et al.: The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) (Version 3). Cancer Imaging Archive (2016). https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK
3. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. (2019)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision **59**(2), 167–181 (2004)
5. Machhale, K., Nandpuru, H.B., Kapur, V., Kosta, L.: MRI brain cancer classification using hybrid classifier (SVM-KNN). In: 2015 International Conference on Industrial Instrumentation and Control (ICIC), pp. 60–65. IEEE, Pune, India (2015). https://doi.org/10.1109/IIC.2015.7150592
6. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
7. Kalavathi, P.: Brain tissue segmentation in MR brain images using multiple Otsu's thresholding technique. In: 2013 8th International Conference on Computer Science & Education (ICCSE), pp. 639–642. IEEE (2013)
8. Glavny Onkologichesky Portal Rossii (Main Oncological Portal of Russia): Official Statistics on Malignant Neoplasms Incidence in 2023. https://glavonco.ru/cancer_register/. Accessed 13 Feb 2025
9. Scikit-image Documentation: Morphological Filtering (2025). https://scikit-image.org/docs/stable/auto_examples/applications/plot_morphology.html. Accessed 25 Mar 2025
10. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 4th edn., pp. 638–648. Publisher, Location (2018)