Clustering based Medical Image Segmentation: A Study on MRI Scans of Brain Tumors

Aigul B. Mimenbayeva*
Departament of Computational and Data Science
Astana IT University
Astana, -, Kazakhstan
Aigulka79_79@mail.ru

Aliya A. Aruova
Department of Computational and Data Science
Astana IT University
Astana, Kazakhstan
a.aruova@astanait.edu.kz

Gulmira K. Bekmagambetova
Department of Computer Science
Agrotechnical Research University named after S. Seifullin
Astana, Kazakhstan
g_bek@mail.ru

Rozamgul S. Niyazova
L.N. Gumilyov Eurasian National University
Department of Artificial Intelligence Technologies
Astana, Kazakhstan
Rozamgul@List.ru

Rakhila D. Turebayeva
L.N. Gumilyov Eurasian National University
Department of Artificial Intelligence Technologies
Astana, Kazakstan
turebayeva_rd@enu.kz

Akgul A. Naizagarayeva
Department of Information Systems
Kazakh Agrotechnical Research University named after S. Seifullin
Astana, Kazakhstan
akgul_1985@mail.ru

Ainur F. Tursumbayeva
Department of Information Systems
Kazakh Agrotechnical Research University named after S. Seifullin
Astana, Kazakhstan
turcumbaewa_ainur84@mail.ru

Abstract

This study investigates the application of Hierarchical clustering for image segmentation, with a focus on its efficacy in analyzing medical images, particularly MRI scans of brain tumors. Image segmentation plays a pivotal role in computer vision, facilitating various applications across industries. Leveraging a systematic approach, we conduct a comprehensive review of recent literature on machine learning algorithms for image segmentation. Subsequently, utilizing a dataset comprising MRI images with and without tumors, we preprocess and analyze the data using the Histogram of Oriented Gradients (HOG) technique to extract pertinent features. These features serve as input for the Hierarchical clustering algorithm to partition the images into distinct regions of interest. For each row of vectors, the Jensen-Shenton distance was calculated. The resulting symmetric matrices are distances among the corresponding vectors, quantifying dissimilarity in cluster analysis. Our findings underscore the effectiveness of Hierarchical clustering in clustering medical images, with potential implications for advancing computational analysis in healthcare and related domains.

*Corresponding author.

ICAAI 2024, London, United Kingdom

ACM ISBN 979-8-4007-1801-4/24/10
https://doi.org/10.1145/3704137.3704174

CCS Concepts

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Unsupervised learning**; • **Cluster analysis.**;

Keywords

Hierarchical clustering, Medical image segmentation, HOG function, MRI Scans, Brain tumors

ACM Reference Format:
Aigul B. Mimenbayeva, Aliya A. Aruova, Gulmira K. Bekmagambetova, Rozamgul S. Niyazova, Rakhila D. Turebayeva, Akgul A. Naizagarayeva, and Ainur F. Tursumbayeva. 2024. Clustering based Medical Image Segmentation: A Study on MRI Scans of Brain Tumors. In *2024 The 8th International Conference on Advances in Artificial Intelligence (ICAAI 2024), October 17–19, 2024, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3704137.3704174

1 Introduction

Computer vision relies heavily on image segmentation technology within computer-aided diagnostic systems. This technology enables precise identification of areas in need of treatment, focusing only on the required characteristics rather than all parts of an image as shown in [1, 2]. By bounding cancerous cells in boxes, image segmentation algorithms can accurately determine the severity of cancer and provide clearer results.

With advancements in computer and internet technologies, medical imaging has become a focal point in modern medicine. Numerous medical images from diverse sources are now available online, providing valuable resources for retrieving medical information. Imaging modalities such as grayscale and color images, MRI scans, thermal, and infrared images [2] offer insights into internal anatomy and bodily functions. Digitalizing these images aids in identifying relevant regions and extracting key features accurately. Therefore, classifying image pixels into anatomical areas (muscular blood vessels) and pathological areas (tissue deformations) is essential [3] for supporting diagnostic and treatment decisions.

This research delves into the application of advanced machine learning techniques to enhance the analysis of medical images, particularly MRI scans of brain tumors. With the increasing significance of computer-aided diagnosis and treatment planning in healthcare, accurate and efficient image segmentation methods are essential for extracting meaningful insights from complex medical data. By leveraging the capabilities of Hierarchical clustering, an unsupervised machine learning algorithm, we aim to delineate regions of interest within the images and facilitate the interpretation of diagnostic information. Traditional methods of manual segmentation are time-consuming and subjective, highlighting the need for automated and reliable techniques as shown in [4].

This research seeks to address these challenges by exploring the application of Hierarchical clustering for medical image segmentation. By demonstrating the effectiveness of this approach in analyzing MRI scans of brain tumors, we aim to contribute to the development of advanced computational tools for healthcare professionals. Ultimately, our research endeavors to improve diagnostic accuracy, treatment outcomes, and patient care in the field of neuroimaging and beyond.

2 Related work

The field of medical diagnostics has seen significant advancements with the emergence of machine learning (ML) techniques, particularly deep learning (DL), in medical imaging analysis. In the context of brain tumor detection, the utilization of ML algorithms presents a promising avenue for improving diagnostic accuracy and aiding in early intervention. This literature review synthesizes findings from various studies that explore the application of ML in detecting brain tumors, highlighting key methodologies, challenges, and future directions.

Rana et al. [5] conducted a systematic literature review on ML and DL applications for disease detection and classification. Their analysis encompassed various modalities for medical imaging and highlighted the efficiency of DL in handling large datasets, enabling accurate disease diagnosis at early stages.

Barragán-Montero et al. [6] discussed the integration of AI methods, including ML, into medical imaging workflows. Their review provided an overview of machine learning methods and their application to various medical imaging tasks, emphasizing the increasing adoption of AI in clinical settings.

Kim et al. [7] provided insights into the history, development, and applications of artificial neural networks (ANNs) in medical imaging. While focusing on the broader applications of ML in healthcare, their review highlighted the transformative potential of DL in computer-aided detection and diagnosis.

The reference [8] focuses on exploring the adoption of techniques such as pre-processing, machine learning and deep learning over the last 15 years and presents a comparative analysis using these techniques. Researchers' previous challenges in tumor detection, as well as future opportunities that can be exploited in their work, are reflected in their discussions. Among the issues covered in the current review articles, emerging clinical problems that are not usually addressed in existing review articles were repeatedly mentioned.

Atmakuri et al. [9] employed deep learning techniques for predicting the phase of brain tumors, showcasing the utility of ML in correlating patients' statuses and predicting future anomalies. Their study demonstrated high accuracy in classifying images of brains with different tumor types, highlighting DL's efficacy in medical imaging analysis.

In [10] study, authors presented a novel stack ensemble transfer learning technique known as "SETL_BMRI", which enables the detection of brain tumors in MRI images with higher accuracy than previously possible, as a new stack ensemble transfer learning model. The SETL_BMRI model uses two pre-trained models, Alex Net and VGG19, to improve its generalization capabilities by training the model itself.

Thirunavukkarasu et al. [11] delved into deep learning-based clinical imaging analysis, discussing automated machine learning platforms and their potential information.

Suganyadevi et al. [12] provided an overview of studies utilizing DL techniques for medical image processing, emphasizing the rapid advancements in AI and its application in various medical specialties, including neurology and oncology. Their review outlined state-of-the-art approaches in medical image analysis, demonstrating the effectiveness of DL in identifying patterns and features in clinical images.

Naizagarayeva et. al [13] Examined deep learning methods and their applicability to electro cardiac signal classification problems. The problem of classification and its specification of ECG is investigated. By leveraging deep learning methods and datasets like PhysioNet, researchers can develop accurate and efficient systems for classifying ECG signals, which can aid in the diagnosis and monitoring of various cardiac conditions.

Idowu et al. [14] emphasized the growing prevalence of brain tumors in African populations and the importance of understanding their etiology for effective prevention strategies. While not directly focusing on ML techniques, this review provides crucial context regarding the epidemiology of brain tumors, underscoring the urgency for improved diagnostic approaches.

Keerthana et al. [15] underscored the significance of early detection of brain tumors, recognizing magnetic resonance imaging (MRI) as a superior modality for detailed and consistent imaging. Their review highlighted methodologies for brain tumor detection, including preprocessing of images, segmentation, feature extraction, clustering, and tumor detection, as pivotal steps in the diagnostic process.

Collectively, these studies underscore the transformative potential of ML techniques, particularly DL, in enhancing brain tumor detection and diagnosis. While significant progress has been made, challenges such as standardization, data quality, and ethical considerations remain pertinent. Future research efforts should focus on addressing these challenges to realize the full potential of ML in improving healthcare outcomes for patients with brain tumors.

3 Methodology

The methodology employed in this study adopts a systematic approach tailored to meet its objectives effectively. Initially, an extensive review of recent literature concerning machine learning algorithms in brain tumor detection, with a particular focus on Hierarchical clustering, is undertaken to acquire pertinent insights and methodologies. Following this, the study harnesses a dataset comprising MRI images depicting both tumor-present and tumor-absent scenarios. These images undergo preprocessing procedures and subsequent analysis to extract meaningful features utilizing the Histogram of Oriented Gradients (HOG) technique [16]. These extracted features serve as input data for the Hierarchical clustering algorithm, facilitating the segmentation of images into discernible regions. The quality of segmentation achieved is subsequently evaluated using the silhouette score metric, ensuring a thorough assessment of the segmentation outcomes.

3.1 Dataset

The dataset utilized in this study focuses on brain tumor imaging analysis and classification. It consists of three main folders: "yes," "no," and "pred," housing a total of 3060 Brain MRI Images. "Yes" folder contains 1500 Brain MRI Images exhibiting tumorous conditions. Contrarily, the "No" folder encompasses 1500 Brain MRI Images showcasing non-tumorous conditions.

This dataset serves as a crucial resource for training and evaluating machine learning models aimed at automated brain tumor detection and classification. Through meticulous analysis of these images, researchers can develop and refine algorithms utilizing deep learning techniques such as Convolutional Neural Networks (CNN) and Transfer Learning (TL). The ultimate goal is to enhance diagnostic accuracy and efficiency in the medical field, particularly in regions where access to expert neurosurgeons may be limited.

3.2 Hierarchical clustering algorithm

Hierarchical clustering is a type of clustering algorithm used to group together similar data points based on their distance or similarity. It builds a hierarchy of clusters, where each data point starts as its own cluster, and then clusters are iteratively merged based on their similarity until all data points belong to a single cluster.

When i = j; and the triangular inequality, then:

$$d\,(i,j) \le d\,(i,k) + d\,(k,j) \quad (1)$$

These (1) properties ensure that the distance measurement is consistent and reliable. There are various distance metrics that can be used for clustering, depending on the nature of the data and the specific problem [17].

Jensen-Shannon Divergence (JSD) can be used as a measure of similarity between probability distributions, including those produced by classifiers. The JSD is a symmetric and smoothed version of the Kullback-Leibler Divergence (KLD), which measures the difference between two probability distributions.

In Figure 1 illustrated of flow chart of Hierarchical clustering algorithm:

Figure 1 outlines a hierarchical clustering algorithm process. It begins with "Start" and proceeds to "List the objects with their features." Next, it calculates the distance matrix using values. Each

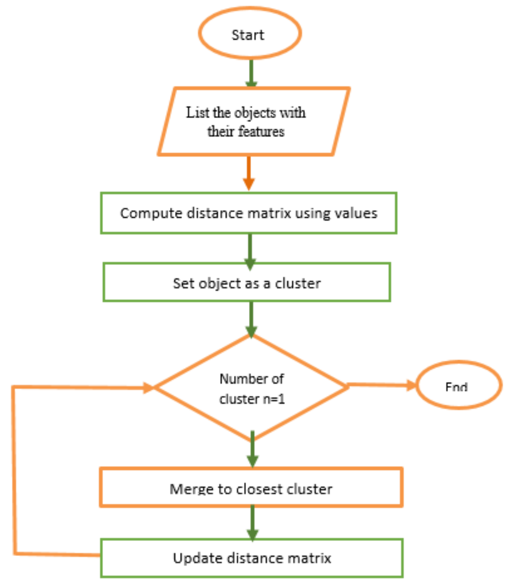

Figure 1: Flow chart of Hierarchical clustering algorithm

object is initially set as a cluster. A decision diamond checks if the number of clusters equals n=1. If false, it merges the two closest clusters and updates the distance matrix, looping back to reevaluate the number of clusters. If true, the process ends. The flowchart uses green and orange elements to indicate different steps and decision points.

Given two probability distributions P and Q, the JSD is computed as follows (2):

$$JSD\,(P,Q) = \frac{1}{2}\left(D_{KL}(P||M) + D_{KL}(Q||M)\right) \quad (2)$$

The JSD ranges from 0 to log2 (for completely dissimilar distributions).

3.3 Silhouette score

The silhouette score is a widely used metric to measure how well data points fit within their clusters. It evaluates both the cohesion (how close a point is to other points within the same cluster) and separation (how distinct the cluster is from other clusters).

The silhouette score $S(i)$ for a data point i is defined as:

$$S\,(i) = \frac{b\,(i) - a\,(i)}{\max\,(a\,(i)\,,\; b\,(i))} \quad (3)$$

Where:
$a(i)$ - average distance of iii to other points in the same cluster (intra-cluster distance).
$b(i)$ - average distance of iii to points in the nearest neighboring cluster (inter-cluster distance).

The score ranges from -1 to 1: 1: Perfectly clustered (point far from other clusters); 0: On or very close to the decision boundary between clusters, -1: Misclassified, closer to a different cluster than its own [18, 19].

3.4 Histogram of Oriented Gradients (HOG) technique

The Histogram of Oriented Gradients (HOG) is a feature descriptor used in computer vision and image processing for object detection and recognition tasks. The computation of HOG involves the following steps:

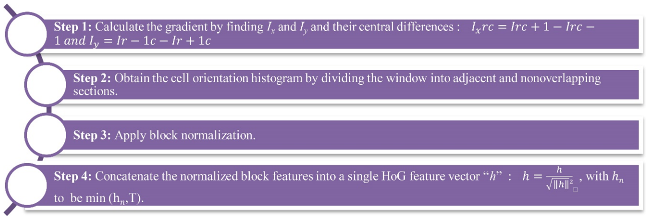

Figure 2: Flow chart of HOG algorithm

Figure 2 shows the block diagram of the HOG algorithm. HOG descriptors are often used as features in machine learning algorithms for tasks such as pedestrian detection, object detection, and face detection.

4 Results & discussion

4.1 Load and Display Images

The proposed methods are executed using a collection of algorithms for image processing created with Python package Scikit-image.

for object detection and recognition tasks. The computation of HOG involves the following steps:

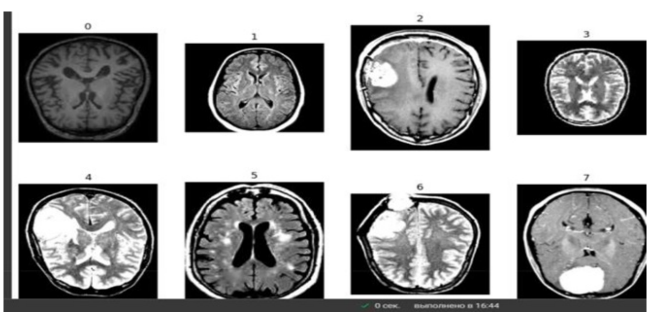

Figure 3: Random images of brain MRI scans

Figure 3 shows a grid of brain MRI scans. These images were loaded into separate lists, images_with_tumor and images_without_tumor. Images with tumors are displayed in the first row of subplots with the title "With tumor", and images without tumors are displayed in the second row with the title "Without tumor" (Figure 3).

4.2 Calculation and Visualization of HOG Features

Histogram of Oriented Gradients (HOG) features were calculated and visualized for both sets of images. Each image with a tumor was gray scale converted and its HOG features were calculated

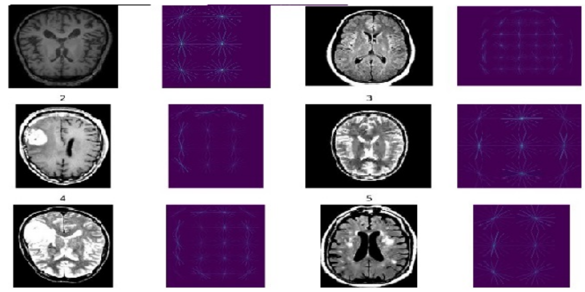

Figure 4: Generated HOG images for each MRI

using a function. The resulting HOG image is shown in the second row of nested plots.

Figure 4 displays brain MRI images in grayscale alongside their corresponding HOG feature visualizations. For each image without tumor, the same operations were performed, and the original images and its corresponding HOG image are displayed in the second row of nested plots (Figure 4).

4.3 Generation of Random Feature Vectors

Next, a synthetic dataset was created by generating random feature vectors. Feature vectors are numerical representations used in machine learning. The number of vectors (n) was set and the fd_list was initialized.

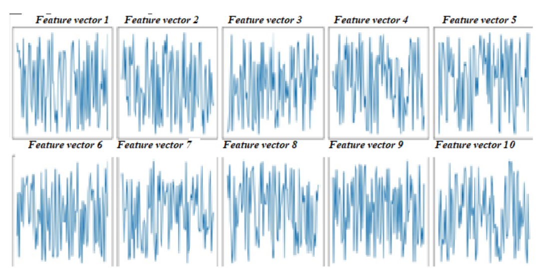

Figure 5: Random Feature Vectors

Figure 5 shows a grid of subplots representing randomly generated feature vectors. Each subplot contains 144 values uniformly distributed between 0 and 1, with indices labeling each feature vector.

Further, it is calculated Jensen-Shannon Distance Matrix:

The Figure 6 illustrates a Jensen-Shannon Distance Matrix. This matrix visually represents the pairwise distances between vectors, computed using the Jensen-Shannon distance metric. It is a symmetric matrix, where smaller values indicate higher similarity between

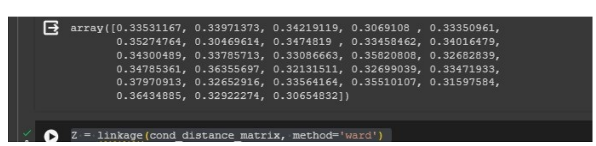

Figure 6: Jensen-Shannon Distance Matrix

vectors. The smallest distance value is 0.306548, representing the minimal dissimilarity between specific clusters.

4.4 Hierarchical Clustering and Dendrogram Visualization

Hierarchical clustering was performed on the Jensen-Shannon distance matrix using the ward method. This method minimizes the variance in cluster formation. Figure 7 shows the resulting dendrogram clearly represents the hierarchical clustering of brain tumors dataset. Each leaf node corresponds to a data point (e.g., photos), and internal nodes represent merged clusters. Branches illustrate clusters formed by the Ward method, with each leaf node corresponding to an individual MRI image or feature vector.

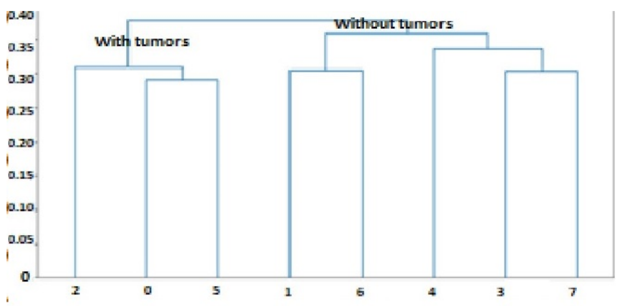

Figure 7: Dendogram Visualization

Images with tumors and images without tumors were processed separately and loaded into lists. These images were plotted using Matplotlib subplots. Figure 8 shows the results: the first row contains images with tumors after segmentation, and the second row shows images without tumors after postprocessing.

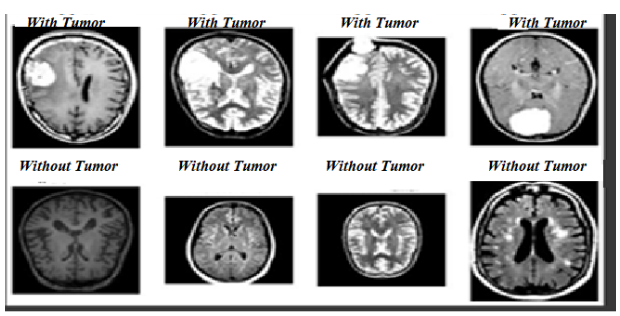

Figure 8: Random images of brain MRIs with and without tumor

The proposed hierarchical clustering algorithm was evaluated using silhouette scores, which averaged 0.85 across all experiments, indicating moderately well-separated clusters. Additionally, it is compared its performance against k-means clustering, which achieved a lower silhouette score 0f of 0.68, demonstrating the effectiveness of hierarchical clustering for brain tumor segmentation. The Jensen-Shannon distance matrix further validated the similarity between clustered images, with a minimum calculated distance of 0.31. The proposed algorithm processed 3,000 images in 120 seconds on average, which, while adequate, suggests room for optimization for larger datasets.

5 CONCLUSION

In this research, a hierarchical clustering algorithm combined with Histogram of Oriented Gradients (HOG) technique was proposed for clustering brain tumor images. Based on a careful analysis of brain MRI images, we demonstrate the effectiveness of hierarchical clustering for accurate image segmentation. Using histogram of oriented gradients (HOG) for feature extraction and Jensen-Shannon distance matrix for cluster estimation, valuable insights into the performance of clustering algorithms in medical image analysis have been uncovered.

Improved data organization and data preprocessing have helped to simplify segmentation, but problems arise in optimizing clustering algorithms to account for parameter variability and sensitivity. These problems underscore the continued need for improved clustering methods in image segmentation tasks. Challenges encountered during the evaluation include parameter sensitivity in the hierarchical clustering algorithm, which required several iterations to achieve optimal results. Future work will explore ensemble learning techniques and hyper parameter optimization to further enhance the segmentation process.

In conclusion, the research contributes to understanding the role of machine learning in medical image analysis, opening up opportunities to improve medical diagnosis and treatment planning.

References

[1] Abdikerimova, G., Yessenova, M., Yerzhanova, A., *et al.* 2023. Applying textural Law's masks to images using machine learning. *International Journal of Electrical and Computer Engineering (IJECE)*, 1, 5 (2023), 5569–5575. DOI: https://doi.org/10.11591/ijece.v13i5.

[2] Mimenbayeva, A. and Zhukabayeva, T. 2020. A review of free resources for processing and analyzing geospatial data. In *Proceedings of the 6th International Conference on Engineering & MIS (ICEMIS '20)*, (2020), 64:1–64:4. DOI: https://doi.org/10.1145/3410352.3410800.

[3] Meiramkhanov, T., Savitsky, B., and Irgaliiyev, E. 2024. Analyzing and Responding to the Spread of COVID-19. In *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2024, 1–5. DOI: https://doi.org/10.1109/SIST61555.2024.10629290.

[4] Amin, J., Sharif, M., Yasmin, M., and Fernandes, S. L. 2018. Big data analysis for brain tumor detection: Deep convolutional neural networks. *Future Generation Computer Systems*, 87 (2018), 290–297. DOI: https://doi.org/10.1016/j.future.2018.04.065.

[5] Rana, M. and Bhushan, M. 2022. Machine learning and deep learning approach for medical image analysis: Diagnosis to detection. *Multimedia Tools and Applications* (2022), 1–39. DOI: https://doi.org/10.1007/s11042-022-14305-w.

[6] Barragán-Montero, A., Javaid, U., Valdés, G., *et al.* 2021. Artificial intelligence and machine learning for medical imaging: A technology review. *Physics in Medicine*, 83 (2021), 242–256. DOI: https://doi.org/10.1016/j.ejmp.2021.04.016.

[7] Kim, M., Yun, J., Cho, Y., *et al.* 2019. Deep Learning in Medical Imaging. *Neurospine*, 16, 4 (2019), 657–668. DOI: https://doi.org/10.14245/ns.1938396.198.

[8] Lotlikar, V. S., Satpute, N., and Gupta, A. 2022. Brain tumor detection using machine learning and deep learning: A review. *Current Medical Imaging*, 18, 6 (2022), 604–622. DOI: https://doi.org/10.2174/1573405617666210923144739.

[9] Atmakuri, M. K., Ram, A. G., and Prasad, V. V. K. D. V. 2023. Reliable image metrics-based brain tumor analysis using sensor deep learning technologies. *Measurement: Sensors*, 29 (2023), 100864. DOI: https://doi.org/10.1016/j.measen.2023.100864.

[10] Natha, S., Laila, U., Gashim, I. A., *et al.* 2024. Automated brain tumor identification in biomedical radiology images: A multi-model ensemble deep learning approach. *Applied Sciences*, 14 (2024), 2210. DOI: https://doi.org/10.3390/app14052210.

[11] Thirunavukarasu, A., Elangovan, K., Gutierrez, L., *et al.* 2023. Democratizing artificial intelligence imaging analysis with automated machine learning: Tutorial. *Journal of Medical Internet Research*, 25 (2023), e49949. DOI: https://doi.org/10.2196/49949.

[12] Suganyadevi, S., Seethalakshmi, V., and Balasamy, K. 2022. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11 (2022), 19–38. DOI: https://doi.org/10.1007/s13735-021-00218-1.

[13] Naizagarayeva, A., Abdikerimova, G., Shaikhanova, A., *et al.* 2023. Detection of heart pathology using deep learning methods. *International Journal of Power Electronics and Drive Systems / International Journal of Electrical and Computer Engineering*, 13, 6 (2023), 6673–6680. DOI: https://doi.org/10.11591/ijece.v13i6.pp6673-6680.

[14] Idowu, O. E. and Idowu, M. A. 2008. Environmental causes of childhood brain tumors. *African Health Sciences*, 8, 1 (2008), 1–4. PMID: 19357723; PMCID: PMC2408544.

[15] Keerthana, A., Kumar, B. K., Akshaya, K. S., and Kamalraj, S. 2021. Brain tumor detection using machine learning algorithm. *Journal of Physics: Conference Series*, 1937, 1 (2021), 012008. DOI: https://doi.org/10.1088/1742-6596/1937/1/012008.

[16] Herholz, K., Langen, K. J., Schiepers, C., and Mountz, J. M. 2012. Brain tumors. *Seminars in Nuclear Medicine*, 42, 6 (2012), 356–370. DOI: https://doi.org/10.1053/j.semnuclmed.2012.06.001.

[17] Mimenbayeva, A., Artykbayev, S., Suleimenova, R., *et al.* 2023. Determination of the number of clusters of normalized vegetation indices using the k-means algorithm. *Eastern-European Journal of Enterprise Technologies*, 2, 125 (2023), 42–55. DOI: https://doi.org/10.15587/1729-4061.2023.290129.

[18] Mimenbayeva, M., Omirtay, R., Niyazova, G., *et al.* 2023. An artificial neural network model to forecast grain crops in North-Kazakhstan Agricultural Experimental Station. In *2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, Plovdiv, Bulgaria, 2023, 1–5. DOI: https://doi.org/10.1109/CIEES58940.2023.10378787.

[19] Yessenova, M., Abdikerimova, G., Baitemirova, N., *et al.* 2022. The applicability of informative textural features for the detection of factors negatively influencing the growth of wheat on aerial images. *Eastern-European Journal of Enterprise Technologies, 4 (2 (118)), 51–58.* DOI: https://doi.org/10.15587/1729-4061.2022.263433