International Journal of Wireless Communications and Mobile Computing
2025, Vol. 12, No. 1, pp. 16-22
https://doi.org/10.11648/j.wcmc.20251201.12

Research Article

Detection of Brain Tumors Using Optimized Features in Clustering Techniques for Enhanced Model Development with Accuracy

Yavanaboina Tezaaw*, Kumba Vijaya Lakshmi

Department of Computer Science, Sri Venkateshwara University, Tirupati, India

Abstract

The growth of cells in the brain or nearby tissues, known as brain tumors, they may be termed benign(non-cancerous) or malignant(cancerous) and can cause various symptoms depending on their location and size. Brain tumor, both benign and malignant cause significant clinical challenges due to their complexity and the diverse range of symptoms they produce depending on their location, size, and type. Tumor classification has traditionally relied on histopathological examination, but molecular insights are becoming crucial in improving accuracy and treatment strategies. Clustering techniques particularly DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identify molecular subtypes in brain tumor datasets, specifically genetic expression data from the GSE50161 dataset. The goal is to improve the detection of Brain Tumor patterns ultimately contributing to better diagnostic, prognostic, and treatment strategies for the patients. Identifying distinct molecular subtypes through genetic expression data can assist in creating personalized treatment plans for patients. By categorizing tumors more accurately, clinicians can choose therapies that target specific molecular mechanisms leading to better outcome Ultimately leading to enhanced accuracy in brain tumor detection.

Keywords

Brain Tumors, Bioinformatics, Gene Expression, Clustering, DBSCAN, Model Predictions

1. Introduction

Clustering algorithms represent unsupervised machine learning technique utilized for grouping similar data points together based on specific criteria. It aims to discover inherent patterns or structures within datasets without requiring labeled data. These algorithms operate by partitioning the data into clusters, wherein data points within the same cluster exhibit greater similarity to each other than to those in other clusters. Common clustering algorithms encompass K-means, hierarchical clustering, and DBSCANK-means clustering involves partitioning the data into k clusters through iterative updates of cluster centroids, aiming to minimize the sum of squared distances between data points and centroids. On the other hand, hierarchical clustering constructs a tree-like structure of clusters by iteratively merging clusters based on their similarity until a stopping criterion is met. DBSCAN clusters data points according to their density, defining clusters as dense regions separated by areas of lower density, while also identifying noise points as outliers. These algorithms find appli-

*Corresponding author: tezaawyavanaboina@gmail.com (Yavanaboina Tezaaw)

Received: 18 February 2025; **Accepted:** 3 March 2025; **Published:** 21 March 2025

cations in various fields, including data mining, pattern recognition, and image segmentation, assisting in tasks such as customer segmentation, anomaly detection, and exploratory data analysis.

The Human Brain controls the body's functions and helps individuals adapt to different situations. It allows people to think, speak, and express emotions [1]. The brain is made up of three main parts: cerebrospinal fluid, white matter, and gray matter. Gray matter regulates brain activity and consists of neurons and glial cells. White matter fibers connect different brain areas, including the cerebral cortex, which is responsible for higher functions, and the corpus callosum, which connects the brain's left and right hemispheres [2].

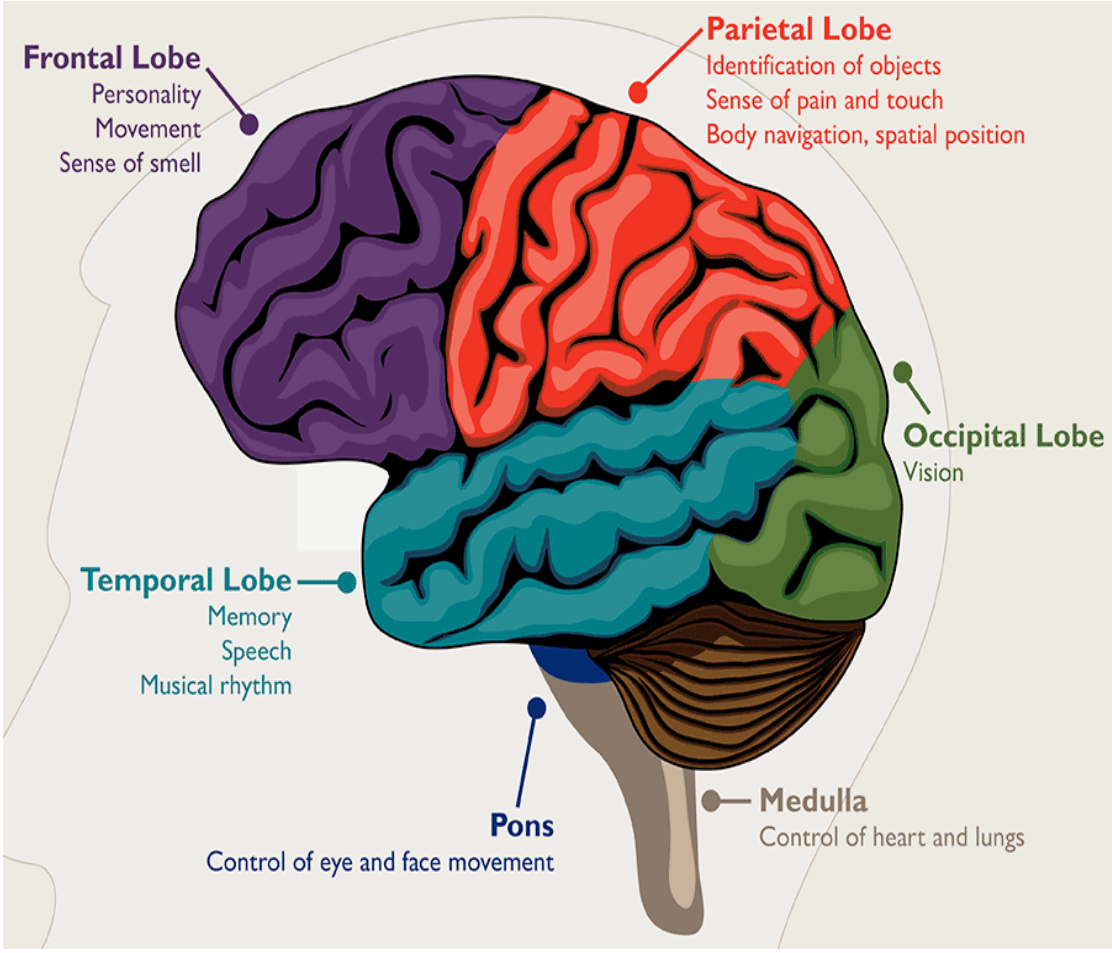

Figure 1. *Overview of brain cerebral functions.*

Brain tumors are complex entities, comprising a diverse spectrum of abnormal cell proliferation within the brain or in the surrounding tissues. These tumors can have significant implications on affecting various neurological functions and leading to various symptoms, spanning from mild cognitive impairments to severe impairments in motor function, sensation, and cognition. The severity of these symptoms can vary depending on factors such as the tumor's location, size, and growth rate. As such, the diagnosis and management of brain tumors require a comprehensive understanding of their molecular and clinical characteristics [3].

In recent times, the field of bioinformatics has become increasingly vital for unraveling the molecular intricacies of brain tumors. At the heart of this endeavor lies the analysis of genetic expression data, which provides a window into the molecular underpinnings of tumor development and progression. By examining the activity levels of genes within tumor cells, researchers can identify patterns and signatures that offer valuable insights into tumor biology [4, 5]. Clustering techniques offer a promising avenue for addressing the challenge of tumor heterogeneity by grouping together samples with similar gene expression profiles. By identifying cohesive clusters representing different tumor subtypes, clustering algorithms enable researchers to reveal concealed patterns and relationships within the population of tumors [6]. This method offers a deeper understanding of tumor biology, laying the groundwork for personalized diagnostic and therapeutic approaches tailored to each patient [7] and summarized with different sections like literature survey in Section 2. Section 3 provides an elaborate discussion on the functions of the pro-

posed system, the dataset description. Section 4 outlines the performance of the proposed approach and presents the results obtained from the experiments. Subsequently, Section 5 comprises the conclusion and outlines avenues for future research with the references.

2. Review of Literature

The review of the literature provides a comprehensive study of the scientific literature and current research to identify different molecular subtypes of brain tumors based on genetic expression data along with prediction models targeting to increase the accuracy.

S. Turgut et al., (2018) [10] concentrated on using several machine learning algorithms to classify patients based on microarray breast cancer data. Without using any feature selection strategies at first, eight different algorithms—SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost, and Gradient Boosting Machines—were used. After that, two different feature selection techniques were used, and the findings were compared with the first classification results as well as with each other. After using the two feature selection techniques, SVM produced the best results out of all the evaluated algorithms. The scientists also investigated how changing the MLP's layer and neuron count affected categorization accuracy. M. M. Mufassirin et al., (2018) [9, 11] used DNA microarray datasets, the authors suggested a unique feature selection method that combines filter and wrapper techniques to improve the accuracy of cancer data categorization. The datasets were first pre-processed using a Gain Ratio Filter using the Ranker search algorithm. They then used a wrapper known as the Wrapper Subset Evaluator using the best initial forward selection searching technique to assess the generated gene subsets. Then, for classification, machine learning classifiers such as Bayes Net, Sequential Minimal Optimization (SMO), Decision Tree (J48), Naïve Bayes, and Deep Learning were used. The accuracy rates obtained from testing on five cancer microarray datasets ranged from 89.69% to 100%, which is promising. Huyen Nguyen et al., (2020) [12] developed hybrid feature selection technique to address the problem of analyzing and categorizing genetic diseases from microarray datasets, which is a difficult and expensive process in biomedical research because of the enormous number of features. This technique effectively selects pertinent genes and saves time by combining the t-test, Fisher ratio, and Bayesian logistic regression. Features were extracted using a modified firefly optimization-based discriminant independent component analysis (MF-DICA), where the modified firefly optimization technique improved search efficiency. Saradhi A. V (2018) [13] presents a successful technique for identifying tumors volume in brain MRI data that uses k-means clustering. The method proposes an automated brain tumors segmentation strategy that combines the k-means clustering for tissue classification with the Perona and Malik anisotropic diffusion model for picture improvement. Malathi M et al (2018) [14] solves the problems of medical image categorization, such as noise and non-uniform texture, are addressed in this paper by presenting an efficient method that combines spatial fuzzy C-means clustering with K-means clustering. By leveraging the advantages of both techniques, the proposed method achieves high accuracy while minimizing computation time. Discrete wavelet transform is utilized for feature extraction, with Back Propagation Algorithm employed for abnormality classification, leveraging extracted features from the MRI brain image. Kumar, D et al., (2021) [13] in their research, four modules—pre-processing, feature extraction, classification, and segmentation—of a machine-learning MRI brain tumour classification system are presented. Initially, to improve classification accuracy, noise is removed from input pictures using the Median Filter. Following the extraction of texture characteristics from pre-processed pictures, an adaptive k-nearest neighbor classifier classifies the collected features as normal or abnormal. Additionally, the best possible fuzzy C-means clustering technique is used to segregate tumor areas. The accuracy, sensitivity, and specificity of the segmentation and classification methods are evaluated using two datasets: the publically accessible dataset and the BRATS MICCAI brain tumor dataset.

Overall, the literature review underscores the critical role of clustering techniques and machine learning algorithms in advancing our understanding of brain tumor biology and improving clinical outcomes through enhanced detection and classification methods.

3. Methodology

Central to our approach is the concept of optimized feature selection, wherein we identify the most informative features (genes) for brain tumor classification [8]. By integrating optimized features into the clustering analysis, our goal is to enhance the accuracy and robustness of our predictive model.

3.1. Data Set

The dataset extracted from Kaggle is utilized for Brain Tumor detection and is stored in a CSV file format. It is represented as a Pandas Data Frame, consisting of 54677 entries. Each entry corresponds to a specific sample or observation. The dataset comprises multiple columns, which likely contain various features extracted from genetic expression profiles or imaging data relevant to brain tumor detection. Additionally, the dataset includes a class column, indicating the type of brain tumor or its absence (labeled as "normal"). There are a total of 131 samples categorized into different classes, with the majority belonging to ependymoma (46 samples), followed by glioblastoma (34 samples), medulloblastoma (22 samples), and pilocytic astrocytoma (15 samples). The "normal" class comprises 13 samples.

3.2. Implementation of Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The operational principle of DBSCAN entails iterating through each data point in the dataset in order to identify clusters based on their density characteristics. Starting from a chosen point 'a', the algorithm retrieves all points that are density-reachable from 'a' with respect to the specified parameters, namely Eps (epsilon) and MinPts (minimum number of points). If 'a' is classified as a core point, meaning it has a sufficient number of neighboring points within the specified radius Eps, a cluster is formed around it. On the other hand, if no points are density-reachable from 'a', it is considered a border point. In this case, the algorithm proceeds to the next point in the dataset. If a point is not in the neighborhood of any other point within the specified Eps, it is labeled as a noisy point or outlier. This process continues iteratively until all points in the dataset have been processed, resulting in the identification of clusters, border points, and outliers. Through this approach, DBSCAN effectively captures the underlying density-based structure of the data, allowing for the discovery of clusters of arbitrary shapes and the detection of outliers.

The mathematical expression for Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm can be represented as follows:

Given a dataset D={x1, x2,..., xn} consisting of n data points in a multidimensional space, DBSCAN aims to partition this dataset into clusters based on density. The algorithm requires two parameters: ε, the maximum radius of the neighborhood around a data point, and minPts, the minimum number of points required to form a dense region.

Core point: xi is considered a core point if there are at least minPts points within a distance ε from it, including itself. Mathematically, this can be expressed as:

$$\text{core}(xi)=\{xj \in D \mid / / xi-xj / / \leq \varepsilon\} \geq \text{minPts}$$

Border point: xi is a border point if it is not a core point but lies within the neighborhood of a core point. Mathematically, this can be expressed as:

$$\text{border}(xi)=\{xi \in D \mid / / xi-xj / / \leq \varepsilon, \text{ where } xj \text{ is a core point}\}$$

Noise point: xi is considered a noise point if it is neither a core point nor a border point.

The clusters are formed by expanding around core points and including border points within the neighborhood. Points that cannot be assigned to any cluster are labeled as noise.

3.3. Algorithm Summary: DBSCAN

Step 1: The algorithm was initialized with the dataset *D* and parameters ε (epsilon) and *minPts* (minimum number of points).

Step 2: For each point *xi* in the dataset, all points within the ε-neighborhood of *xi* were searched, denoted as $N\varepsilon(xi)$.

Step 3: Core points were identified by determining if a point *xi* qualified as a core point based on the condition: $|N\varepsilon(xi)| \geq minPts$.

Step 4: Clusters were formed for each core point *xi*. If *xi* hadn't been assigned to a cluster, a new cluster *C* was created. The cluster was recursively expanded by adding all reachable points within the ε neighborhood to *C*.

Step 5: Border points, which were points within the ε neighborhood of a core point but weren't core themselves, were assigned to their respective clusters.

Step 6: Noise points were identified as unassigned points, those not assigned to any cluster.

Its capability to identify clusters of arbitrary shapes and handle noise and outliers in the dataset makes DBSCAN suitable for various applications, including spatial data analysis, anomaly detection, and image segmentation.

4. Result and Discussion

In this study, dataset collection involved gathering relevant data pertaining to brain tumor patterns, including genetic expression profiles from a curated source. This includes the use of a label encoder to transform categorical variables into numerical format, enabling compatibility with machine learning algorithms. Additionally, robust scaling was applied to normalize the numerical features, ensuring uniformity in scale and reducing the impact of outliers. Subsequently, the DBSCAN clustering technique was employed to cluster the original features based on their density distribution. By combining the original features with the cluster labels generated by DBSCAN, a new dataset was constructed, providing enhanced insights into the inherent structure of the data. Figure 2, Clustering plot illustrates the cluster labels ranging from 0.95 to 1.

Model evaluation was conducted in a two-phase approach, encompassing both training and testing phases. The dataset underwent division into training and testing sets, with the test phase comprising 30% of the data.

Different classifier models were utilized for evaluation, in which top 5 models are including LGBM, XGB, Extra Trees, Gradient Boosting, Random Forest. Each model was trained on the training dataset using the original features combined with cluster labels generated by DBSCAN. Performance metrics, such as accuracy, precision, recall, and F1-score, were employed to evaluate the effectiveness of each model in detecting brain tumors. The Extra Trees classifier emerged as the best model in this study, attaining an accuracy of 0.974 during evaluation.

Table 1. Comparion of Accuracy, Balanced Accuracy, Precision, Recall, F1 score Support.

Models	Accuracy	Balanced Accuracy	Precision	Recall	F1 score	Support	Time
Extra tree	o.974	0.985	1	0.93	0.96	14	0.2
Random forest	0.948	0.935	1	0.93	0.96	14	1.08
LGBM	0.948	0.965	0.93	0.93	0.93	14	24.63
Bagging	0.948	0.942	1	1	1	14	5.96
Decision tree	0.923	0.945	0.93	0.93	0.93	14	1.06

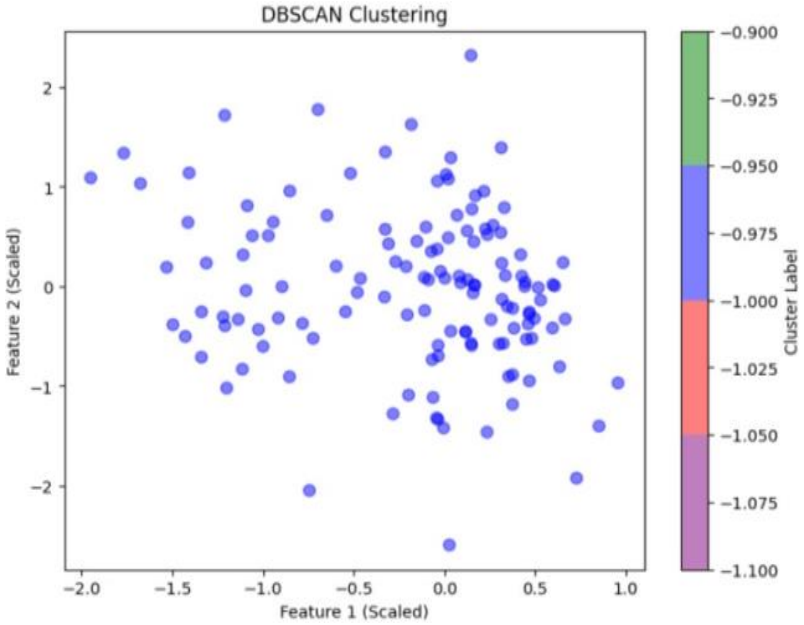

Figure 2. Clustering plot.

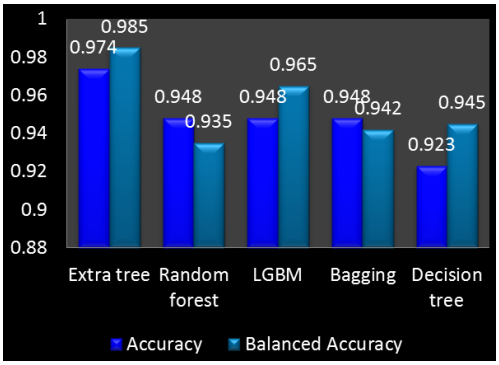

Figure 3. Accuracy and Balanced Accuracy for considered top5 ML Models.

In the Figure 3 chart the accuracy and balanced accuracy of all the compared models. Among them, the Extra Trees classifier stands out with the highest accuracy obtained. The Extra Trees classifier achieved an impressive accuracy of 0.974.

In the Figure 4, Extra Tree classifier also exhibited efficient computational performance, completing the task within a relatively short time period of 0.26 seconds. This combination of high accuracy and fast processing time underscores the effectiveness and practical utility of the Extra Trees classifier in real-world applications, particularly in the domain of brain tumor detection.

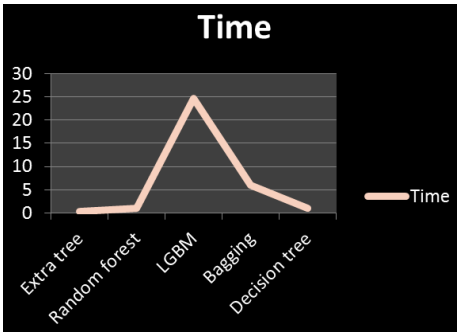

Figure 4. Time consumed for considered for top 5 models ML models.

In further optimization efforts, the Optuna framework was employed to fine-tune hyperparameters for enhanced model performance. A total of 50 trials were conducted to search for the best combination of hyperparameters. Through systematic exploration of the hyperparameter space, Optuna identified the optimal set of hyper parameters that maximized the model's accuracy. This iterative tuning process allowed for the refinement of model parameters, ultimately improving the accuracy of the classifier.

The confusion heat matrices visually represent how well the tuned and untuned models classify different classes of brain tumor patterns. By comparing these matrices, we can observe any improvements or changes in the model's ability to correctly identify and distinguish between different types of brain tumors after tuning the hyper parameters is represented in following Figure 5.

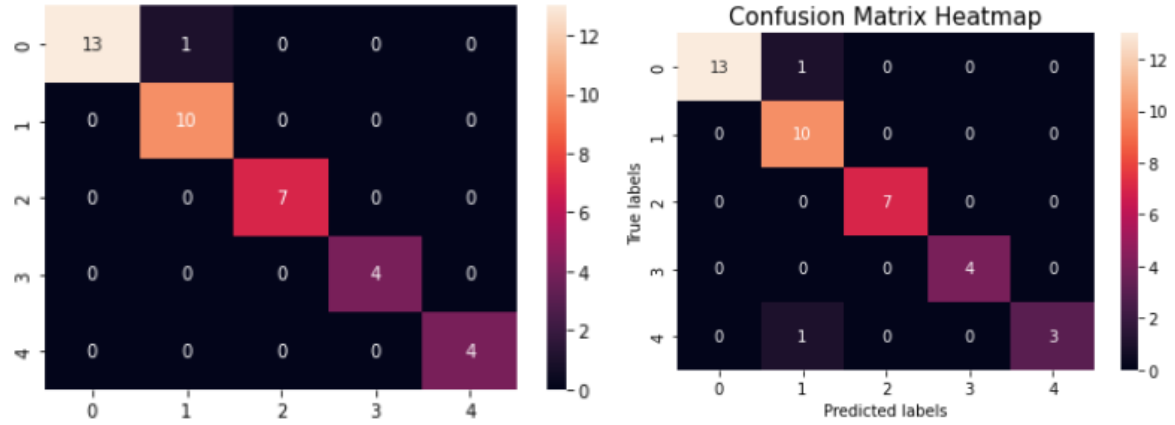

Figure 5. *Confusion matrix metrics for tuned and untuned Extra tree lassifiers.*

The Figure 6 depicts a comparison of the accuracy and balanced accuracy scores between the tuned and untuned models in precisely categorizing brain tumor patterns while accounting for class imbalances in the dataset.

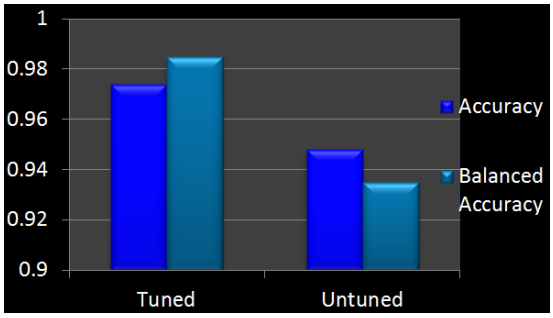

Figure 6. *Accuracy and Balanced Accuracy for Tuned and Untuned Extra Classifiers.*

5. Conclusion and Future Work

In conclusion, the implementation and evaluations of our system demonstrate superior detection performance compared to other contemporary systems. The optimized tuned system achieved a Brain tumor detection accuracy of 97.4%, surpassing that of other modern systems. Moving forward, future work will focus on expanding the dataset to ensure robustness across various scenarios. Additionally, employing advanced techniques such as grid search for parameter optimization and increasing computational resources will further enhance the system's performance. These improvements aim to enhance the accuracy and efficiency of brain tumor detection, contributing to advancements in medical imaging technology.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Amin J., Sharif M., Raza M., Saba T., Anjum M. A. Brain tumor detection using statistical and machine learning method. Comput. Methods Programs Biomed. 2019; 177: 69–79. https://doi.org/10.1016/j.cmpb.2019.05.015

[2] Khan A. H., Abbas S., Khan M. A., Farooq U., Khan W. A., Siddiqui S. Y., Ahmad A. Intelligent model for brain tumor identification using deep learning. Appl. Comput. Intell. Soft Comput. 2022; 2022: 8104054. https://doi.org/10.1155/2022/8104054

[3] Abdel-Gawad AH, Lobna A, Ahmed S, Radwan G (2020) Optimized Edge Detection Technique for Brain Tumor Detection in MR Images. IEEE Access 8: 136243–136259.

[4] B. R. Reddy, Y. Vijay Kumar and M. Prabhakar, "Clustering large amounts of healthcare datasets using fuzzy c-means algorithm," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 93-97.

[5] X. Li, Y. Kang, Y. Zhu, G. Zheng and J. Wang, "An improved medical image segmentation algorithm based on clustering techniques," 2017 10th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI), Shanghai, 2017, pp. 1-5.

[6] Budati A. K., Katta R. B. An automated brain tumor detection and classification from MRI images using machine learning technique s with IoT. Environ. Dev. Sustain. 2022; 24: 10570–10584. https://doi.org/10.1007/s10668-021-01861-8

[7] Murtagh, Fionn& Contreras, Pedro. (2017). Algorithms for hierarchical clustering: An overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 7. https://doi.org/10.1002/widm.1219

[8] Cilia ND, De Stefano C, Fontanella F, Raimondo S, Scotto di Freca A. An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets. Information. 2019; 10(3): 109. https://doi.org/10.3390/info10030109

[9] M. M. Mufassirin, R. G. Ragel, A novel filter-wrapper based feature selection approach for cancer data classification, 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2018, pp. 1-6.

[10] Huyen Nguyen, T. T., Van, P. N., Tran, Q., XuanVo, N., & Quang Vo, T. (2020). Cancer classification from microarray data for genomic disorder research using optimal discriminant independent component analysis and kernel extreme learning machine. International Journal for Numerical Methods in Biomedical Engineering, 36.

[11] Saradhi, A. V. (2018). An Efficient Method K Detection of Tumour Volume in Brain MRI Scans.

[12] M M, P S. MRI Brain Tumour Segmentation Using Hybrid Clustering and Classification by Back Propagation Algorithm. Asian Pac J Cancer Prev. 2018 Nov 29; 19(11): 3257-3263. https://doi.org/0.31557/APJCP.2018.19.11.3257

[13] Kumar, D. M., Satyanarayana, D. & Prasad, M. N. G. MRI brain tumor detection using optimal possibilistic fuzzy C-means clustering algorithm and adaptive k-nearest neighbor classifier. J Ambient Intell Human Comput 12, 2867–2880 (2021).

[14] Zhang, S., Zeng, T., Hu, B., Zhang, Y. H., Feng, K., Chen, L., et al. (2020). Discriminating Origin Tissues of Tumor Cell Lines by Methylation Signatures and Dys-Methylated Rules. Front. Bioeng. Biotechnol. 8, 507.