Pertanika J. Sci. & Technol. 33 (3): 1105 - 1117 (2025)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Clustering-based Hybrid Approach for Analyzing High-grade Gliomas Using Radiomic Features

Sairam Vuppala Adithya[1], Navaneeth Bhaskar[2]* and Priyanka Tupe-Waghmare[1]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), 412115 Pune, India
[2]Department of Artificial Intelligence & Data Science, NMAM Institute of Technology, NITTE (Deemed to be University), Nitte - 574110, Karnataka, India

ABSTRACT

Unlabeled data is a significant problem in healthcare and other fields that deal with huge datasets. Unsupervised learning has the potential to be an effective solution in this case. The use of unsupervised algorithms in disease diagnosis has not been widely explored. In this work, we have developed a clustering algorithm to analyze the gliomas using Magnetic Resonance Imaging (MRI) data. Glioma is a severe medical illness that necessitates an accurate and timely diagnosis to establish effective treatment options. We used Pyradiomics to extract radiomic characteristics from MRI scans, which were then fed into a number of clustering methods, with cluster fitness assessed using primary assessment metrics. The best clustering algorithm was used as the pre-processor and to train major classification algorithms. In this study, we examined the performance of three prominent clustering algorithms, with agglomerative clustering outperforming the others. We achieved 0.83 Silhouette Coefficient, 0.21 Davies-Bouldin Index, and 323.22 Calinski-Harabasz Index values using aggregative clustering using Pyradiomics features. The decision tree strategy outperformed all classification methods, achieving 99.54% accuracy when clustering was applied to preprocess the data before classification. The proposed work has considerable potential for faster and more accurate analysis of medical image problems, especially in gliomas.

Keywords: Classification Algorithms, clustering, gliomas, machine learning, magnetic resonance imaging, unsupervised learning

ARTICLE INFO

Article history:
Received: 10 February 2024
Accepted: 16 January 2025
Published: 26 March 2025

DOI: https://doi.org/10.47836/pjst.33.3.01

E-mail addresses:
sairam.v.mtech2023@sitpune.edu.in (Sairam Vuppala Adithya)
navbskr@gmail.com (Navaneeth Bhaskar)
priyanka.tupe@sitpune.edu.in (Priyanka Tupe-Waghmare)
* Corresponding author

INTRODUCTION

A glioma is a tumor that develops from glial cells, the central nervous system's support cells. Glioma is deemed dangerous for various reasons, including its nature and influence on the central nervous system. Gliomas are graded according to the criteria

specified by the World Health Organization (WHO). Grades 1 and 2 are considered low-grade gliomas, grade 3 is astrocytoma, and grade 4 is classified as glioblastoma (Chen et al., 2017). Grade 3 and grade 4 are classified as the aggressive category. A biopsy is the standard approach to diagnosing glioma. It is a surgical procedure that involves taking the sample from the affected area for pathological examination. Analysis using biopsy provides accurate results. However, it has certain drawbacks, such as proliferation, the need for mastery, time obligation, and the possibility of more cancer growth. Magnetic Resonance Imaging (MRI) is a non-invasive, non-surgical, quick imaging technology and a viable alternative for glioma analysis (Tupe-Waghmare et al., 2021).

The status of Isocitrate Dehydrogenase (IDH), the codeletion status of 1p19q, and the methylation status of O6-methylguanine-DNA-methyltransferase (MGMT) are the major molecular features for the diagnosis of gliomas. Monitoring the status of these biomarkers is important for evaluating the glioma profile and anticipating prognosis (Zheng et al., 2020). However, advanced techniques and medical procedures are required to analyze the status of these biomarkers using MRI images. An alternative diagnosis approach is to use data-driven decision-making using artificial intelligence. Artificial intelligence and machine learning algorithms are efficient and successful tools that assist physicians and support them in making decisions, boosting their confidence in making accurate diagnoses.

Unsupervised learning approaches are commonly used where data labeling is impossible or required. It is mainly applied to identify correlations in data. Labeling the data requires a high level of clinical skills in healthcare applications. Thus, it can be costly and time-consuming (Dike et al., 2018). Clustering is mainly adopted when a novel disease is most likely to happen with no prior medical records. Unlike supervised learning algorithms, where we train the model based on labeled data with already recorded values, unsupervised learning uses unlabeled data to identify unique patterns, relationships, or clusters. The unavailability of labeled data in clustering algorithms becomes a substantial barrier to evaluating the quality and validity of the clusters formed. Unsupervised clustering methods such as K-Means, Fuzzy clustering, Hierarchical clustering, and Kernel K-Means are commonly implemented; however, the model's efficiency is dependent on their ability to validate results (Govender & Sivakumar, 2020).

Naeem et al. (2023) have analyzed unsupervised learning techniques, including the Apriori algorithm, frequent pattern growth algorithm, k-means clustering, and principal component analysis. The clustering approaches are mainly categorized into two types: hierarchical clustering and partition clustering. The applications of unsupervised learning in domains such as machine vision, speech recognition, self-driving cars, and natural language processing have been highlighted in this review paper. Like machine learning approaches, unsupervised techniques do not require labeled data or manual feature selection, which causes flexibility and automation issues. Supervised learning techniques are used in most

healthcare research involving automated disease detection. Despite these, unsupervised learning techniques remain underexplored, especially in applications like disease prediction. In recent work, Mansour et al. (2021) developed an innovative unsupervised model called a variational autoencoder to predict COVID-19 cases. They trained and tested the model through several experiments, demonstrating its impressive performance, achieving high accuracy rates of 98.7% for binary classification and 99.2% for multi-class classification.

Bibi et al. (2022) have shown that unsupervised learning can be a good option for solving classification problems. Their research used concept-based and hierarchical clustering methods to analyze Twitter sentiments. The authors have combined popular hierarchical clustering techniques like single linkage, complete linkage, and average linkage in sequence. The authors have shown that unsupervised learning techniques, like supervised learning methods, can perform well. In a study by Zhang et al. (2023), hyperspectral imaging (HSI) and unsupervised classification techniques were used to identify normal and necrotic areas in small intestinal tissues. K-means and density peaks (DP) clustering algorithms were utilized to distinguish between these tissue types. Their results showed that the DP clustering algorithm attained an average clustering purity of 92.07%. They have concluded that HSI, along with DP clustering, can assist doctors in identifying normal and necrotic tissue in the small intestine.

Bhattacharjee et al. (2022) proposed unsupervised learning approaches to distinguish between benign and malignant stages of the prostate gland using images. Radiomic characteristics have been extracted from the entire slide image using important clustering techniques such as spectral clustering, agglomerative clustering, K-means, K-medoids, and the Gaussian mixed model (GMM). These methods were assessed using Silhouette and Rand scores. It was found that the best results were generated for the K-means algorithm. Song et al. (2023) used Chest X-rays to detect pneumonia. They used and analyzed unsupervised learning techniques for the detection. The X-ray pictures were transformed to grayscale and scaled to consistent dimensions. The radiometric characteristics were retrieved. Two algorithms were used for clustering, i.e., k-means clustering and spectral clustering. For spectral clustering, the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index values were 0.44, 1.025, and 311.5, respectively, while for k-means clustering, the values were 0, 8 and 0.28.

Unsupervised learning techniques study glioma analysis, focusing mainly on brain tumor segmentation (Bougacha et al., 2018). In many research studies, unsupervised learning techniques are used as part of their pipeline to segment the Region of Interest (RoI) and then supervised learning techniques are applied to improve the results. According to our current analysis, only a countable number of research have implemented unsupervised learning for preprocessing to cluster and identify subgroups in gliomas. The proposed research aims to provide a clustering technique for automatically detecting the status of

IDH biomarker index and grade of gliomas from MRI data. This is done by extracting radiomic features and using preprocessed data to improve the classification algorithm's performance. This generic framework can cluster patients based on common features, which can be highly advantageous for disease analysis. A public glioma dataset is used to test and verify the proposed methodology. The proposed method can be applied to any medical imaging resource for disease analysis, including those without labels.

MATERIALS AND METHODS

Our work used the Cancer Genome Atlas (TCGA) dataset (Ganini et al., 2021). TCGA has genetic, clinical, and imaging data related to glioblastoma. MRI scans of patients, information on grades, and other biomarkers are present in the dataset. This study inspected 210 3D volumes of grades 3 and 4 with IDH mutation status using T1-weighted, Fluid-Attenuated Inversion Recovery (FLAIR), and T2-weighted modalities (Al-Saeed et al., 2009). Figure 1 portrays the workflow for the proposed method. Statistical features are repossessed from the generated 3D volumes and used to create clustering algorithms. Trimming the image eliminates unwanted sections and emphasizes areas of interest. Here, several slices from the volume that do not contain gliomas have been deleted as we are interested mainly in gliomas. Based on this objective, the volumes were condensed and normalized to have voxel values ranging from 0 to 1. In our work, three main clustering algorithms: K-means clustering, agglomerative clustering, and Balanced Iterative Reducing and Clustering (BIRCH) (Wahyuningrum et al., 2021; Madan & Dana, 2016) are implemented. These models' performance is evaluated by assessing various performance metric values.

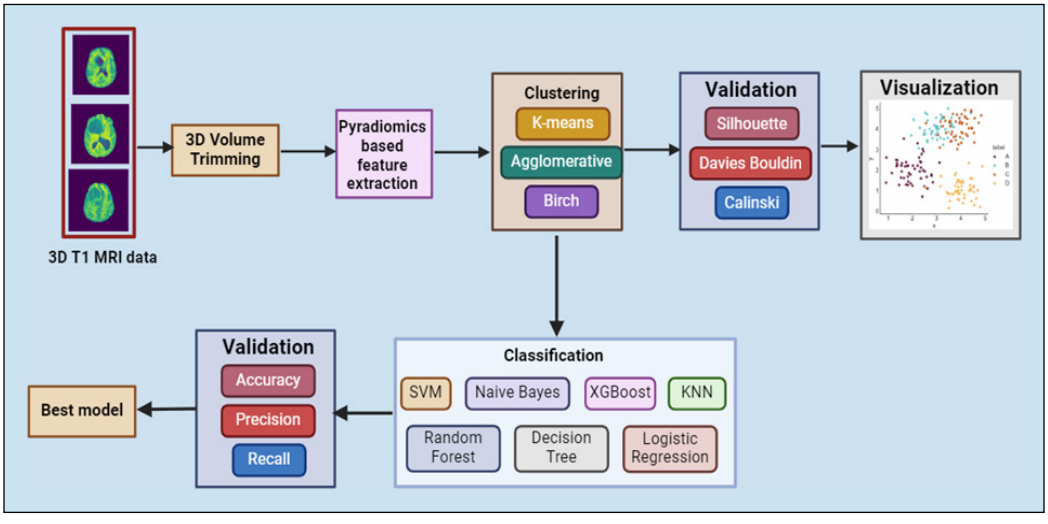


Figure 1. Workflow for the proposed approach

Feature Extraction and Reduction

The Pyradiomics module was used for automated feature extraction (Van Griethuysen et al., 2017). Pyradiomics is the Python package commonly used to extract features from images. The automated feature extractor tool automatically computes and extracts relevant statistical features from the data. This tool uses the image and the mask as input for feature extraction. The model has extracted 120 features from each medical image and mask data using statistics, shape, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM).

The features extracted are First Order Statistics (19), Shape-2D (16), Shape-3D (10), GLCM (24), GLRLM (16), GLSZM (16), NGTDM (5), and GLDM (14). The generated dataset has 120 columns, which is extremely complex. Processing such a huge number of columns is time-consuming and can be expensive. This can result in postponed training of the clustering algorithms. This issue is addressed by using a suitable dimensionality reduction technique. Dimensionality reduction has to be performed before implementing the clustering algorithm. This work utilized the Principal Component Analysis (PCA) technique to minimize the dimensions (Jolliffe & Cadima, 2016). It is the best technique to deal with the challenges when a greater number of features are available per specimen.

Clustering Models

Three major clustering algorithms, K-means clustering, agglomerative clustering, and BIRCH, have been implemented and compared in our study. K-means is a well-known unsupervised clustering algorithm. It is the simplest technique, but it can handle complex data sets. Initially, centroids equivalent to a number of clusters are formed. Then, the distance between each data point and all the centroids is calculated. Next, the centroids are figured once again with the previously computed distance. This process is repeated until the centroids converge and remain constant (Wahyuningrum et al., 2021). Agglomerative clustering is a hierarchical and unsupervised learning method. A bottom-up approach is used in this process. Link distances are calculated based on the resemblance between two data points. Every data point is part of a single cluster, and the nearby points are combined using distance-based linkages. This method is repeated until all data points have been combined into a single cluster (Griffiths et al., 1984). BIRCH is another hierarchical clustering procedure. If the data set contains multiple features and is also huge, even then, this dynamic clustering method performs well. This method clusters data points using a height-balanced methodology and a feature tree (Madan & Dana, 2016).

Classification Models

We have used seven different machine-learning classification algorithms in this study. These include Support Vector Machine (SVM), Naive Bayes, KNN (K-Nearest Neighbor), Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost) (Bhaskar, Bairagi et al., 2023). Logistic regression is a simple classification algorithm, and it is mainly used for binary classification problems. It uses the most common sigmoid activation function on the input data and classifies the output based on a certain limit.

Support Vector Machine is a commonly used machine learning technique as it is highly flexible. In Naive Bayes, the conditional probability of each attribute is examined separately using the Bayesian theorem principle. K-Nearest Neighbor is a learning approach that does not rely on labeled data, and it attempts to map a new data point to its nearest neighbor and perform the grouping accordingly. Decision trees and random forests are tree-based models. They make use of entropies and information gain to make predictions. XGBoost is another ensemble learning method that uses boosting to gain greater accuracy (Choudhary et al., 2022; Bhaskar, Tupe-Waghmare et al., 2023).

RESULTS AND DISCUSSION

The features extracted from the MRI images are converted to a CSV data file with a dimension of 70,128, with 128 features extracted from each of the 70 MRI volumes. We have performed clustering using the three unsupervised algorithms considered in this study. Three major performance parameters, the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index values (Ashari et al., 2023), were used to evaluate the performance of these models. The Silhouette Coefficient is calculated using the following Equation 1:

$$S(i) = (b(i) - a(i)) / max\{a(i), b(i)\} \quad [1]$$

where S(i) is the Silhouette Coefficient for data point i, a(i) represents the intra-cluster distance, and b(i) represents the inter-cluster distance.

Davies-Bouldin matrix index is used to measure the cluster fitness. This method can be used to assess the suitability of various data divisions. The goal is to bring this index as near to zero as possible. The following Equation 2 represents the formula for the Davies-Bouldin index:

$$DB = (1 / k) * \sum(i = 1 \; to \; k) \; max(R(ij)), where \; i \neq j \quad [2]$$

Where k is the number of clusters, and R(ij) is the measure of dissimilarity between cluster I and cluster j.

The Calinski-Harabasz Index is theoretically determined as the inter-cluster and intra-cluster dispersion ratio. A higher value for this index indicates that the observations within each cluster are dense and well-separated (Aik et al., 2023). Calinski-Harabasz Index value can be determined with the following Equation 3:

$$CH = (B / (k - 1)) / (W / (n - k)) \quad [3]$$

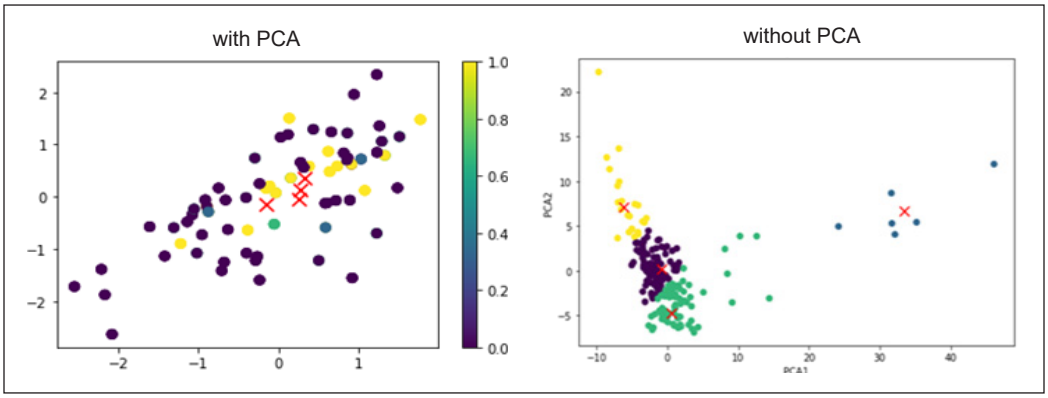
Where B is the variation between clusters, k is the number of clusters, W is the variance inside the clusters, and n is the total number of data points.

Table 1 displays the results of various clustering algorithms, both with and without PCA feature reduction. The Agglomerative clustering algorithm outperformed all other algorithms in our analysis, producing the highest Silhouette Coefficient and Calinski-Harabasz Index scores and the lowest Davies-Bouldin Index scores. PCA has significantly enhanced the performance of unsupervised algorithms, with scores increasing almost double when compared to algorithms trained on the same dataset without PCA. Figure 2 shows scatterplot representations of K-means and BIRCH clustering techniques with and without PCA. The scatterplots produced show the separation of clusters generated by various clustering approaches. Each point on the plot represents a data point, and the values of the features define its location. The scatter plot generated for the agglomerative clustering algorithm is shown in Figure 3. The plot clearly distinguishes between categories, indicating that the clustering method effectively groups comparable data points.

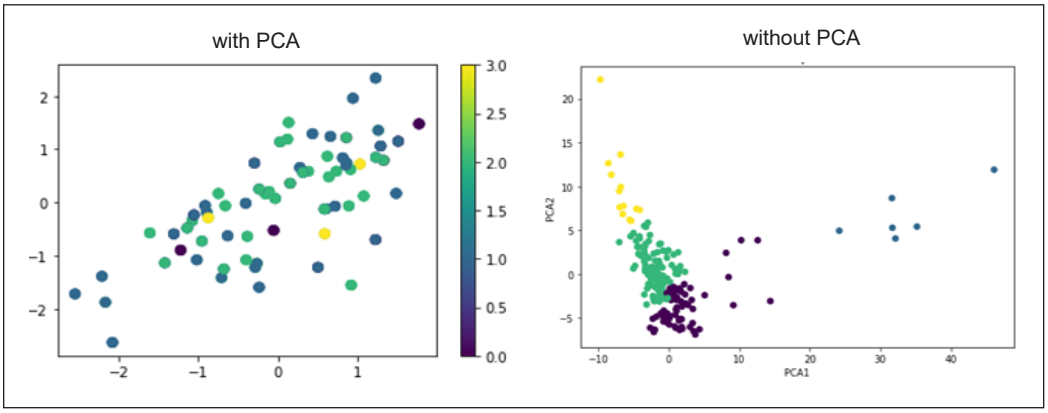
The dendrogram diagram of the agglomerative clustering showing the hierarchical relationships between different entities is depicted in Figure 4. Entities refer to the individual data points clustered by the agglomerative clustering technique. Each branch of the dendrogram symbolizes the merging or splitting of clusters, and the length of the branches reflects the clusters' dissimilarity or distance. The dendrogram illustrates the graphical depiction of the clustering results that allow for a better understanding of the links between distinct data points or clusters.

Table 1
Performance metrics obtained for different models compared in this study

Clustering Algorithm	Feature Reduction	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
K-means	No	0.399	1.21	78.06
	PCA	0.529	0.70	300.9
Agglomerative	No	0.77	0.37	69.5
	PCA	0.83	0.21	323.22
BIRCH	No	0.242	1.51	70.2
	PCA	0.55	0.64	265.05



(a)



(b)

Figure 2. Scatterplot representations (a) K-Means clustering with and without PCA (b) BIRCH clustering with and without PCA

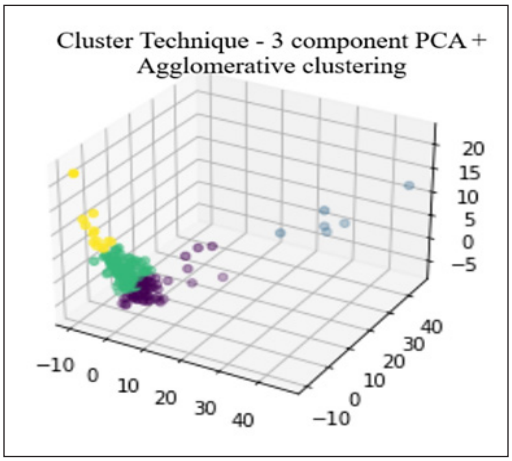


Figure 3. Scatterplot representations for agglomerative clustering with PCA

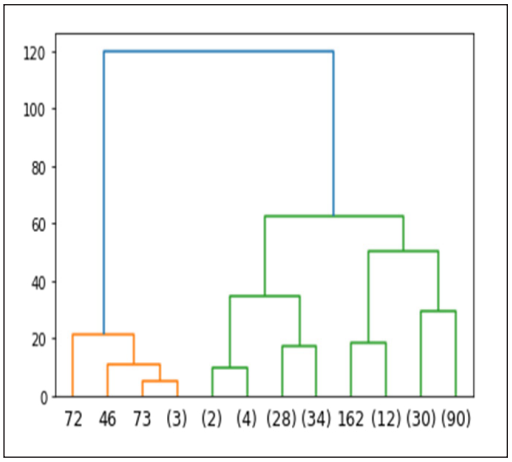


Figure 4. The dendrogram diagram for the agglomerative clustering algorithm

Clusters are produced more effectively using PCA. In high-dimensional spaces, the distance between points can become meaningless, and data points can be sparsely distributed. PCA decreases dimensionality by focusing on the most relevant features, resulting in more meaningful distances and clearly defined clusters. As seen in Figure 3, all clustering algorithms struggled to discriminate between the green and dark blue color groups. These clusters potentially correspond to data points from the G3 mutant and G4 wild classes, which have been clinically shown to be significantly related. Additional clustering approaches can help provide a more complete picture of the data's intrinsic structures. Combining results from different clustering algorithms using ensemble approaches may improve overall clustering performance. Integrating autoencoders into the process can also help to increase feature learning.

Out of the three clustering approaches examined in this study, agglomerative clustering produced the best results and was thus selected as the pre-processor. The labels acquired from agglomerative clustering are used to train classification algorithms. We measured the primary performance parameters to evaluate the models' performance. Table 2 shows the supervised classification algorithms' accuracy, precision, and recall values on the training and testing sets after applying the clustering technique. Figure 5 shows a visual comparison of the validation accuracy of the training models.

It is apparent that using agglomerative clustering as a preprocessing step improves the performance of classification systems. We observed an average improvement of nearly 3 percentage points in accuracy after incorporating agglomerative clustering as a preprocessing step. By structuring data into meaningful clusters before training, we observed accuracy, precision, and recall metrics improvements across various classification algorithms. This approach optimizes model training and facilitates better pattern recognition and predictive accuracy in complex datasets. The SVM and Naive Bayes models achieved less than 90% validation accuracies, while the other models produced more than 90% validation accuracies. The validation accuracy of 99% was achieved for the decision tree

Table 2
Performance evaluation of machine learning techniques with clustering

Models	Training Data			Validation		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
SVM	80.6	93.75	81.11	76.19	92.64	82.14
Naive Bayes	92.72	88.57	96.06	83.33	71.38	90.47
KNN	97.57	98.78	92.23	95.23	97.17	97.17
XGBoost	100	100	100	95.23	97.17	97.17
Logistic Regression	100	100	100	97.61	98.33	98.95
Decision Tree	100	100	100	99.54	99.33	99.33
Random Forest	100	100	100	99.34	99.12	99.12

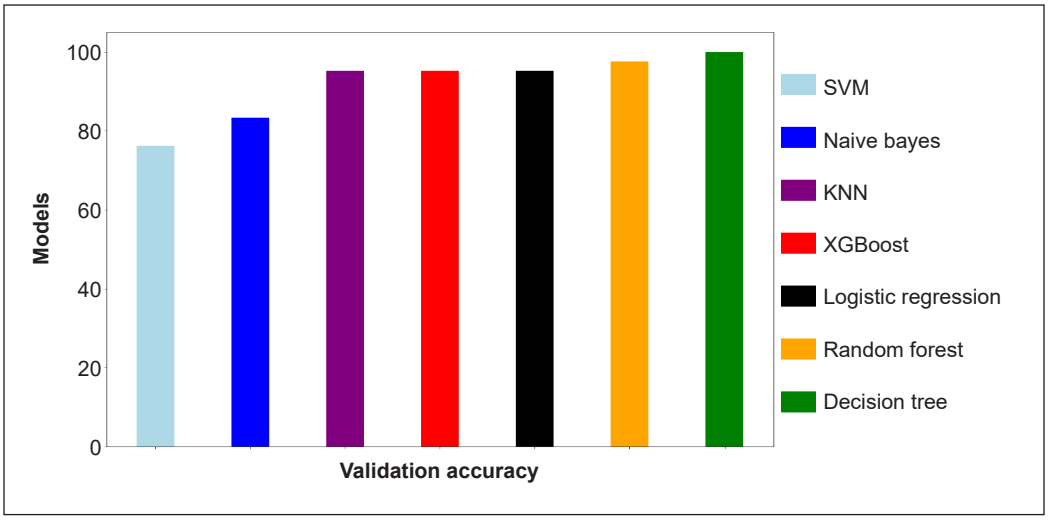


Figure 5. Comparison of the validation accuracy of the training models with clustering

and random forest. In addition, the precision and recall scores are very high, which justifies that the models were properly trained. The models have not shown signs of overfitting, with just a negligible variation in training and testing performance. The clustering approach grouped the grade and IDH based on similarities, which may have assisted the machine learning models in better training, thus enhancing the performance. Furthermore, reducing the dimensions using PCA before clustering also improved the performance of all three techniques, with a 7% increase in silhouette coefficient and a 16% decrease in the Davies-Bouldin score.

It is clearly noted that machine learning algorithms trained on clustered datasets consistently performed better than those on datasets without clusters in terms of accuracy, precision, and recall. Models like XGBoost, decision tree, and random forest exhibited momentous overfitting on the unclustered dataset, which was eased when clustering was applied. These implementations have supported the fact that, by integrating clustering techniques in the preprocessing phase, the performance of the machine learning algorithms is significantly uplifted in identifying grade and mutation status. This method not only decreases the need for extensive dataset labeling but also streamlines the process of anomaly detection in medical imaging applications. By integrating dimensionality reduction techniques, we aim to increase efficiency by not compromising performance, making this method a better choice for handling large, unlabeled datasets in healthcare applications.

The unsupervised learning method proposed in this paper can significantly support clinical practitioners, reducing the time and cost associated with diagnosing oncology patients. This technique requires minimum labeling and thus allows clinicians to process

and analyze large datasets of medical images effortlessly. Applying clustering techniques to new patient data enables automatic annotation, which can then train supervised classifiers for detecting abnormalities in new cases. Incorporating dimensionality reduction techniques further improves algorithm efficiency and speed without reducing the performance. Overall, this approach provides a powerful, time-saving solution for handling unlabeled data, streamlining diagnostic workflows, and aiding in treatment planning in oncology.

CONCLUSION

In this paper, we have used statistical feature extraction techniques and classification algorithms. Clustering of MRI images for glioma analysis using unsupervised algorithms is presented. Accuracy and reliability are ensured by evaluating the results. The performance of classification algorithms is significantly improved by using the proposed framework. These machine learning algorithms have yielded results comparable to deep learning techniques, with the advantage of lower computational costs. The agglomerative clustering algorithm outperformed all other algorithms in our analysis, with the highest Silhouette Coefficient and Calinski-Harabasz Index scores and the lowest Davies-Bouldin Index scores. The decision tree-based model outperformed all classification approaches in the testing set, achieving an accuracy of 99.54% when clustering was used as a preprocessing step. The results show that the proposed system may be used for quick and accurate glioma analysis without requiring computationally intensive hardware.

ACKNOWLEDGEMENT

We thank NMAM Institute of Technology, NITTE (Deemed to be University), Karnataka, India, and Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, for their support and facilities in conducting this research.

REFERENCES

Aik, L. E., Choon, T. W., & Abu, M. S. (2023). K-means algorithm based on flower pollination algorithm and Calinski-Harabasz Index. In *Journal of Physics: Conference Series* (Vol. 2643, No. 1, p. 012019). IOP Publishing. https://doi.org/10.1088/1742-6596/2643/1/012019

Al-Saeed, O., Ismail, M., Athyal, R. P., Rudwan, M., & Khafajee, S. (2009). T1-weighted fluid-attenuated inversion recovery and T1-weighted fast spin-echo contrast-enhanced imaging: A comparison in 20 patients with brain lesions. *Journal of Medical Imaging and Radiation Oncology, 53*(4), 366-372. https://doi.org/10.1111/j.1754-9485.2009.02093.x

Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index evaluation on K-means algorithm for classifying flood-affected areas in Jakarta. *Journal of Applied Informatics and Computing, 7*(1), 95-103. https://doi.org/10.30871/jaic.v7i1.4947

Bhaskar, N., Bairagi, V., Boonchieng, E., & Munot, M. V. (2023). Automated detection of diabetes from exhaled human breath using deep hybrid architecture. *IEEE Access, 11*, 51712-51722. https://doi.org/10.1109/ACCESS.2023.3278278

Bhaskar, N., Tupe-Waghmare, P., Nikam, S. S., & Khedkar, R. (2023). Computer-aided automated detection of kidney disease using supervised learning technique. *International Journal of Electrical and Computer Engineering (IJECE), 13*(5), 5932-5941. https://doi.org/10.11591/ijece.v13i5.pp5932-5941

Bhattacharjee, S., Hwang, Y. B., Sumon, R. I., Rahman, H., Hyeon, D. W., Moon, D., Carole, K. S., Kim, H. C., & Choi, H. K. (2022). Cluster analysis: Unsupervised classification for identifying benign and malignant tumors on whole slide image of prostate cancer. In *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)* (pp. 1-5). IEEE Publishing. https://doi.org/10.1109/IPAS55744.2022.10052952

Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., & Gadekallu, T. R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters, 158*, 80-86. https://doi.org/10.1016/j.patrec.2022.04.004

Bougacha, A., Boughariou, J., Slima, M. B., Hamida, A. B., Mahfoudh, K. B., Kammoun, O., & Mhiri, C. (2018). Comparative study of supervised and unsupervised classification methods: Application to automatic MRI glioma brain tumors segmentation. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (pp. 1-5). IEEE Publishing. https://doi.org/10.1109/ATSIP.2018.8364463

Chen, R., Smith-Cohn, M., Cohen, A. L., & Colman, H. (2017). Glioma subclassifications and their clinical significance. *Neurotherapeutics, 14*, 284-297. https://doi.org/10.1007/s13311-017-0519-x

Choudhary, S., Kumar, A., & Choudhary, S. (2022). Prediction and comparison of diabetes with logistic regression, Naïve Bayes, random forest, and support vector machine. In *International Conference on Innovations in Computer Science and Engineering* (pp. 273-283). Springer. https://doi.org/10.1007/978-981-19-7455-7_20

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)* (pp. 322-327). IEEE Publishing. https://doi.org/10.1109/CBS.2018.8612259

Ganini, C., Amelio, I., Bertolo, R., Bove, P., Buonomo, O. C., Candi, E., Cipriani, C., Daniele, N. D., Juhl, H., Mauriello, A., Marani, C., Marshall, J., Melino, S., Marchetti, P., Montanaro, M., Natale, M. E., Novelli, F., Palmieri, G., Piacentini, M., ... & Melino, G. (2021). Global mapping of cancers: The cancer genome atlas and beyond. *Molecular Oncology, 15*(11), 2823-2840. https://doi.org/10.1002/1878-0261.13056

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research, 11*(1), 40-56. https://doi.org/10.1016/j.apr.2019.09.009

Griffiths, A., Robinson, L. A., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*(3), 175-205. https://doi.org/10.1108/eb026764

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374*(2065), Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Madan, S., & Dana, K. J. (2016). Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering. *Pattern Analysis and Applications, 19*, 1023-1040. https://doi.org/10.1007/s10044-015-0472-4

Mansour, R. F., Escorcia-Gutierrez, J., Gamarra, M., Gupta, D., Castillo, O., & Kumar, S. (2021). Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. *Pattern Recognition Letters, 151*, 267-274. https://doi.org/10.1016/j.patrec.2021.08.018

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithm: Comprehensive review. *International Journal of Computing and Digital Systems, 13*(1), 911-921. http://dx.doi.org/10.12785/ijcds/130172

Song, J., Gu, Y., & Kumar, E. (2023). Chest disease image classification based on spectral clustering algorithm. *Research Reports on Computer Science, 2*(1), 77-90. https://doi.org/10.37256/rrcs.2120232742

Tupe-Waghmare, P., Malpure, P., Kotecha, K., Beniwal, M., Santosh, V., Saini, J., & Ingalhalikar, M. (2021). Comprehensive genomic subtyping of glioma using semi-supervised multi-task deep learning on multimodal MRI. *IEEE Access, 9*, 167900-167910. https://doi.org/10.1109/ACCESS.2021.3136293

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J. C., Pieper, S., & Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research, 77*(21), e104-e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

Wahyuningrum, T., Khomsah, S., Suyanto, S., Meliana, S., Yunanto, P. E., & Al Maki, W. F. (2021). Improving clustering method performance using K-means, mini batch K-means, BIRCH and spectral. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 206-210). IEEE Publishing. https://doi.org/10.1109/ISRITI54043.2021.9702823

Zhang, L., Huang, D., Chen, X., Zhu, L., Xie, Z., Chen, X., Cui, G., Zhou, Y., Huang, G., & Shi, W. (2023). Discrimination between normal and necrotic small intestinal tissue using hyperspectral imaging and unsupervised classification. *Journal of Biophotonics, 16*(7), Article 202300020. https://doi.org/10.1002/jbio.202300020

Zheng, L., Zhang, M., Hou, J., Gong, J., Nie, L., Chen, X., Zhou, Q., & Chen, N. (2020). High-grade gliomas with isocitrate dehydrogenase wild-type and 1p/19q codeleted: A typical molecular phenotype and current challenges in molecular diagnosis. *Neuropathology, 40*(6), 599-605. https://doi.org/10.1111/neup.12672