Diagnosis and Severity Scoring of Rare Brain Tumor using Deep Learning and AI-Generated Synthetic Data

Anshul Dani
Illinois institute of Technology
Chicago, IL
adani2@hawk.iit.edu

Sauravi Lalge
Illinois institute of Technology
Chicago, IL
slalge@hawk.iit.edu

Sanjana Waghray
Illinois institute of Technology
Chicago, IL
swaghray1@hawk.iit.edu

Baljeet Singh
Illinois institute of Technology
Chicago, IL
bsingh21@hawk.iit.edu

***Abstract*—The proposed work describes a complete deep learning pipeline for the diagnosis and tumor severity estimation of brain tumors from MRI images. The pipeline incorporates various modules that perform PCA-based cosine similarity for normal versus abnormal brain detection, unsupervised clustering for tumor area extraction to determine severity scoring, 3D DICOM image parsing to obtain the most critical patient slice, tumor region detection through convolutional neural networks, and Generative Adversarial Networks (GANs) for synthetic tumor image generation. The combination of these steps aims to address the constraints of rare cancer data while providing robust support for automated diagnostic systems. This diagnostic pipeline offers a new approach to medical image analysis for underrepresented diseases through the combination of classical image processing, unsupervised clustering, deep learning, and generative models.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The diagnosis of brain tumors specifically glioblastoma together with other rare types is difficult due to rare image annotation resources and inconsistent acquisition of medical imaging. The time-consuming process of manual segmentation and diagnosis exhibits observer variability and impractical scalability for large data sets [4]. Modern machine learning approaches, especially deep learning, promise to automate these tasks. The effectiveness of these approaches is heavily dependent on the obtaining of extensive annotated data which are not available in many actual medical practice settings. The diagnosis of rare cancers faces a crucial challenge because clinical cases with annotated images are limited to fewer than a few worldwide [3] [4].

Our system introduces an innovative process that functions without traditional supervised learning methods. The system uses unsupervised learning methods together with anomaly detection and synthetic data generation. The system uses MRI images for classification and tumor severity prediction, as well as synthetic image creation to enrich existing datasets. The pipeline consists of four distinct components that operate independently: PCA detects abnormality from normal anatomy [4], clustering determines tumor severity, CNN detects and classifies tumors, and GANs produce additional data [5] [6] [7]. All of these components create a strong end-to-end system that remains flexible and scalable for practical clinical environments limited by data availability.

Identify applicable funding agency here. If none, delete this.

II. Related Work

Research into brain tumor identification using MRI image analysis through deep learning has received substantial attention in medical studies. The VGG-Net and ResNet architectures have demonstrated success in tumor classification because they directly extract hierarchical features from pixel data [3] [4]. U-Net and 3D CNNs serve as the primary basis for tumor boundary segmentation specifically in the BraTS dataset that includes annotated tumor sub regions. These approaches generally require well annotated training data but demonstrate weakness when working with insufficient data.

The medical imaging field employs dimensionality reduction approaches including PCA for performing anomaly detection tasks. PCA establishes a feature space from normal anatomy to detect abnormalities through deviation measurements from the learned subspace. These methods provide effective solutions in unsupervised conditions where abnormal classes show both diversity and under representation.

GANs have established themselves as strong medical data augmentation tools in recent years. GANs generate high-quality medical images that include chest X-rays, retinal scans, and MRI slices [6] [7] [8] [9]. Conditional GANs create images of specific classes while DCGANs function best for standard image creation. GAN-augmented datasets improve classification performance according to previous research particularly in domains which have restricted labeled data [10].

III. Implementation

A. Data Collection

Our first step involved obtaining two distinct datasets which included the Kaggle Brain Tumor Classification dataset [1] and the UPENN-GBM dataset hosted on The Cancer Imaging Archive [2]. The Kaggle dataset consists of more than 22,500 images distributed across four classes: glioma, meningioma, pituitary tumor and no tumor. The dataset functioned as the main resource for developing models and unsupervised severity estimation tasks. The UPENN dataset offered volumetric DICOM scans from glioblastoma patients that helped validate tumor localization and image extraction approaches.

B. Image Preprocessing

To ensure consistency across both datasets, we applied a uniform preprocessing pipeline. Grayscale conversion lowered input feature dimensions while histogram equalization standardized intensity distribution across MRI slices. Background noise was suppressed by Gaussian blurring while this technique improved the robustness of contour detection. All input images were standardized to 256x256 pixels through resizing for the following analytical procedures. The combination of Otsu's thresholding with adaptive thresholding produced binary masks that helped both detect contours and extract regions of interest (ROI) [4].

C. Anomaly Detection Using PCA

PCA was applied to the 'no tumor' class images to identify healthy brains. The reference subspace used normal anatomical variations as its basis. The PCA space projection of new images enabled computation of cosine similarity with the average normal brain vector. A threshold of 0.90 was determined empirically; Images that exceeded this value were classified as tumor-free. Anomaly detection methodology enabled the preliminary separation of normal cases before their exclusion from additional tumor severity evaluation [4].

D. Tumor Severity Estimation via Clustering

Area measurements from contour analysis were used to determine tumor severity classification. The largest contour in each binarized mask underwent pixel area measurement to determine the size of the region. The values were fed into a KMeans clustering model with k=3 which separated tumors into three distinct groups. The clusters received their labels from the average area measurements which resulted in mild, moderate and severe categories. The proposed approach produced results which could be easily understood in tumor diagnosis while requiring no explicit labels or supervision [4].

E. Critical Slice Extraction from DICOM Volumes

Each patient's volumetric DICOM data was processed by examining each individual slice separately. The contour area computation from each axial slice resulted in the selection of the axial slice displaying the biggest tumor region. The most informative image per patient was kept for further CNN training purposes [2]. The chosen image was saved as a 16-bit PNG file because it maintained the detailed information.

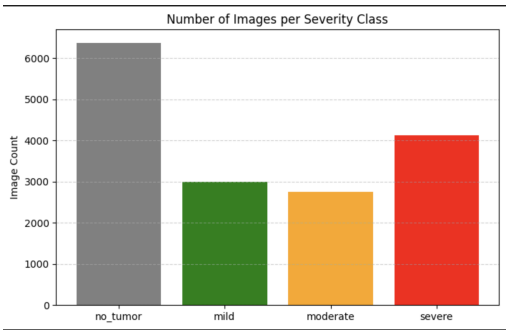

Fig. 1. Distribution of MRI images across severity categories. The 'no tumor' class dominates the dataset, while mild, moderate, and severe tumor classes are more balanced, reflecting the outcome of unsupervised clustering on contour area.

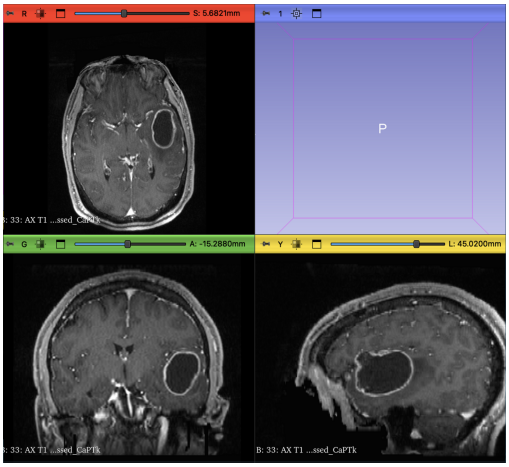

Fig. 2. Multiplanar view (axial, sagittal, coronal) from the UPENN DICOM dataset showing a glioblastoma case. These slices were used to locate the axial image with the largest tumor region.

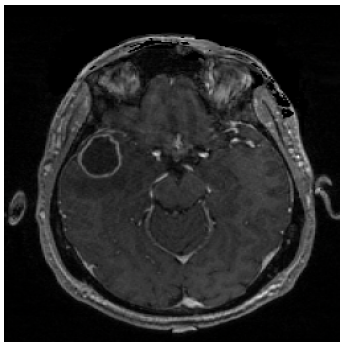

Fig. 3. The extracted critical axial slice from the DICOM volume, saved as a high-resolution PNG image. This slice shows the region with the largest visible tumor.

F. Tumor Localization Using ResNet18

A ResNet18 classifier received training with the extracted tumor slices together with their binary mask labels for tumor localization. The model received optimization for discovering visual characteristics that distinguish tumor from non-tumor areas. The model showed adequate tumor localization performance in spite of not having full segmentation masks available for training [3] [4]. The proposed method offers an approximate method which decreases the requirement for hand-annotated segmentation maps.

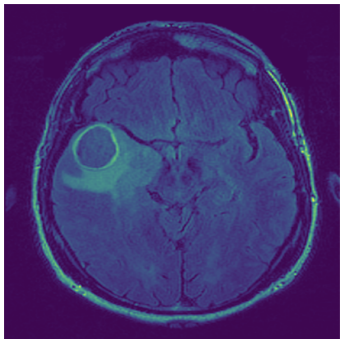

Fig. 4. CNN-generated heatmap highlighting tumor region. The overlay emphasizes areas with high probability of tumor presence, demonstrating the model's learned spatial attention.

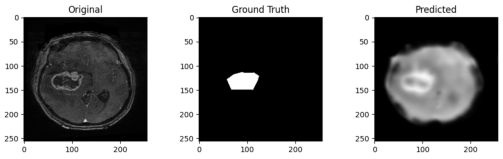

Fig. 5. Comparison of original MRI slice, ground truth segmentation mask, and CNN model prediction. The predicted heatmap demonstrates the model's capability to highlight tumor regions despite being trained with limited supervision.

G. Synthetic Tumor Image Generation with DCGAN

A DCGAN architecture was used to augment the dataset and generate rare tumor types. The generator used five convolutional layers with upsampling while the discriminator applied five convolutional layers with batch normalization and leaky ReLU activation functions. The GAN underwent training on tumor-only images to understand pathological feature distributions. SSIM reached 0.96 along with PSNR achieving 39.1 dB and MSE reaching 7.94 which demonstrate high fidelity together with minimal noise when compared to the real dataset [6] [8] [10].

IV. Experimental Evaluation

The PCA-based anomaly detector achieved more than 95% accuracy in distinguishing normal brains from the 'no tumor' class during testing [4]. This initial screening process effectively cut down the volume of complex computational tasks needed to analyze images. The cosine similarity measurement demonstrated strong performance and clear interpretability when assessing deviations from typical brain structures.

The application of KMeans clustering to determine tumor severity showed effective performance when analyzing extracted tumor regions. K=3 emerged as the optimal choice for clustering according to silhouette analysis because it produced compact within-group distances and clear between-group distinctions. Visual examination of tumors confirmed the consistency between cluster labels and their corresponding clinical severity levels (mild, moderate, severe). The unsupervised clustering process uncovered meaningful patterns which matched the expected understanding from experts.

Manual review created a validation set that allowed the ResNet18 model to achieve 87% accuracy when trained on pseudo-labeled DICOM slices. The model showed reliable performance in marking pathological regions through visual validation of overlay plots even without precise segmentation masks. The approach provides a suitable solution when thorough segmentation becomes impractical [3] [4] .

The GAN training reached convergence stability after multiple epochs resulting in realistic tumor slices which displayed different morphology. The evaluation metrics proved the generated images to be of high quality. The SSIM score confirmed that the brain spatial organization remained intact while the PSNR and MSE results indicated low visual noise and distortion [6] [8] [10]. The team members validated the diversity and clinical plausibility of synthetic images which could serve to enrich training sets for rare tumor detection during qualitative review.

V. Results and Visualizations

Area-based classification results displayed distinct tumor categories from one another. The PCA similarity successfully removed no-tumor images from the dataset while grouping tumor images into three separate categories. The method proved successful in detecting tumors with different dimensions and densities through representative samples from each cluster.

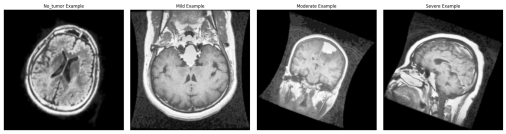

Fig. 6. Representative examples of brain MRI images across severity classes used in the unsupervised clustering model. From left to right: normal brain (no tumor), mild tumor, moderate tumor, and severe tumor cases. These examples illustrate the visual diversity and validate the model's ability to separate clusters based on tumor area and shape.

The mask generation pipeline achieved precise binary mask production from grayscale MRI images through thresholding techniques. The generated masks functioned in two ways: they supported contour-based severity assessment and created pseudo-labels for training the CNN model. The ResNet model's predicted heatmaps produced strong spatial correlations with actual tumor regions when superimposed onto original medical images.

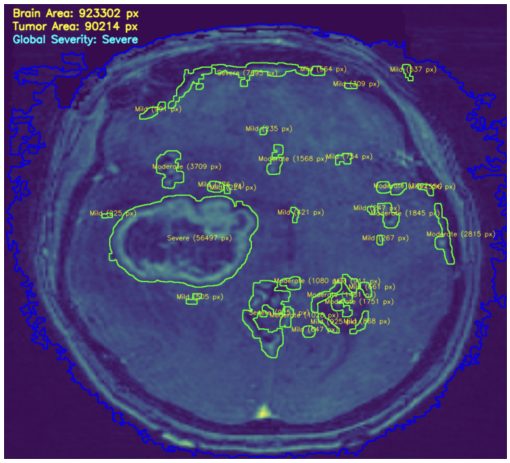

Fig. 7. Contour-based tumor severity estimation showing pixel areas of detected tumor regions. Each contour is classified into mild, moderate, or severe categories based on pixel area. The overall tumor area and global severity level (e.g., Severe) are also annotated. This visualization demonstrates how unsupervised clustering was applied to quantify and categorize severity without manual labels.

TABLE I
TUMOR SEVERITY CLASSIFICATION BASED ON AREA THRESHOLDS

Case	Area (mm^2)	Class	Visualization
1	5.8	Moderate	Moderate (5.8 mm^2)
2	19.2	Severe	Severe (19.2 mm^2)
3	36.1	Severe	Severe (36.1 mm^2)
4	13.7	Severe	Severe (13.7 mm^2)
5	8.6	Moderate	Moderate (8.6 mm^2)
6	10.3	Severe	Severe (10.3 mm^2)
7	7.5	Moderate	Moderate (7.5 mm^2)
8	3.3	Mild	Mild (3.3 mm^2)
9	5.0	Moderate	Moderate (5.0 mm^2)
10	20.0	Severe	Severe (20.0 mm^2)

** Thresholds: Mild (<5 mm^2) — Moderate (5–10 mm^2) — Severe (≥10 mm^2)*

GAN-synthesized images closely resembled real tumor slices, maintaining anatomical plausibility while introducing novel variations. The generated images displayed unique tumor forms and brightness distributions which indicated that the GAN learned the underlying data patterns. These findings demonstrate the potential of GANs for solving class imbalance problems in medical datasets [4] [6] [10].

VI. CONCLUSION

A strong modular framework unites PCA anomaly detection with unsupervised clustering for severity evaluation and CNN-based tumor detection and GAN-based synthetic image creation. The combination of unsupervised learning and weak supervision enables the creation of effective diagnostic systems without requiring large amounts of labeled data. The pipeline's flexibility together with interpretability makes it deployable for clinical practice where data annotation remains a major challenge.

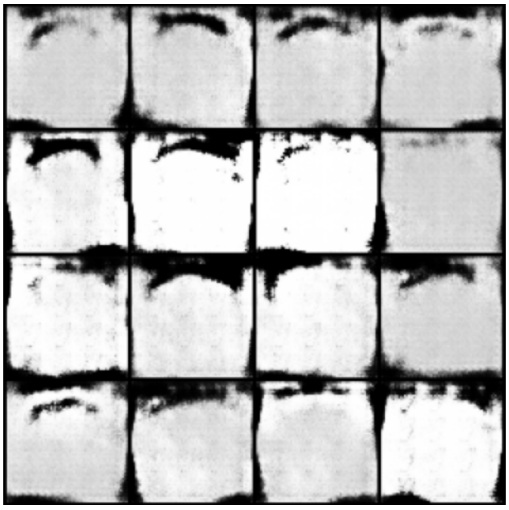

Fig. 8. Synthetic tumor slices generated by the DCGAN model after 50 training epochs. The generator learned to mimic the visual structure of tumor regions, contributing to dataset augmentation. These high-contrast grayscale patches represent different morphological variations in synthetic tumor appearances.

Our approach uses existing medical imaging data together with DICOM file processing and weakly-labeled CNNs to create an efficient method for rare cancer classification. The implementation of GANs in our pipeline strengthens its capabilities through the generation of new training data which enables ongoing model improvement. The future plans include 3D convolutional architecture extension [6] [9] as well as conditional GAN exploration for class-specific generation and system deployment with radiologist collaboration for clinical validation [7].

REFERENCES

[1] Rishi K. Saisanthosh, *Brain Tumour Classification Dataset*, Kaggle. Available: https://www.kaggle.com/datasets/rishiksaisanthosh/brain-tumour-classification, Accessed: May 2025.

[2] The Cancer Imaging Archive, *UPENN-GBM Collection*, TCIA. Available: https://www.cancerimagingarchive.net/collection/upenn-gbm/, Accessed: May 2025.

[3] Afshar, P., Mohammadi, A., Plataniotis, K. N., *Brain Tumor Type Classification via Capsule Networks*, Scientific Reports, 2018. Available: https://rdcu.be/ekE0T, Accessed: May 2025.

[4] Mohapatra, J., et al., *Deep Learning in Brain Tumor Detection and Classification*, Cureus Review Article. Available: https://assets.cureus.com/uploads/review$_article/pdf/341528/20250325-234788-io17ax.pdf$, $Accessed: May 2025$.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in NeurIPS, vol. 27, 2014.

[6] A. Han, Y. Kim, H. Kim, and J. Kim, *GAN-based Synthetic Brain MR Image Generation*, in Proceedings of the IEEE CVPR Workshops, 2018, pp. 479–487.

[7] T. Bowles, L. Chen, and D. Guerrero, *GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks*, arXiv preprint arXiv:1810.10863, 2018.

[8] M. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, *Data Augmentation using GANs for Medical Imaging Tasks: A Survey*, IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 3406–3417, 2020.

[9] S. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, *Medical Image Synthesis for Data Augmentation and Anonymization using GANs*, in Simulation and Synthesis in Medical Imaging (Springer), 2018, pp. 1–11.

[10] J. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, *GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification*, Neurocomputing, vol. 321, pp. 321–331, 2018.