Detection of Isocitrate Dehydrogenase Mutated Glioblastomas Through Anomaly Detection Analytics

Birra Taha, MD[‡]
Taihui Li, MS[§]
Daniel Boley, PhD[§]
Clark C. Chen, MD, PhD[‡*]
Ju Sun, PhD[§*]

[‡]Department of Neurosurgery, University of Minnesota Medical School, Minneapolis, Minnesota, USA; [§]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, USA

*Clark C. Chen and Ju Sun contributed equally to this work.

Correspondence:
Clark C. Chen, MD, PhD,
Department of Neurosurgery,
University of Minnesota Medical School,
500 SE Harvard St,
Minneapolis, MN 55455, USA.
Email: cchen@umn.edu

Ju Sun, PhD,
Department of Computer Science and Engineering,
University of Minnesota,
200 Union St SE,
Minneapolis, MN 55455, USA.
Email: jusun@umn.edu

Received, August 29, 2020.
Accepted, February 16, 2021.

BACKGROUND: The rarity of Isocitrate Dehydrogenase mutated (mIDH) glioblastomas relative to wild-type IDH glioblastomas, as well as their distinct tumor physiology, effectively render them "outliers". Specialized tools are needed to identify these outliers.

OBJECTIVE: To carefully craft and apply anomaly detection methods to identify mIDH glioblastoma based on radiomic features derived from magnetic resonance imaging.

METHODS: T1-post gadolinium images for 188 patients and 138 patients were downloaded from The Cancer Imaging Archive's (TCIA) The Cancer Genome Atlas (TCGA) glioblastoma collection, and from the University of Minnesota Medical Center (UMMC), respectively. Anomaly detection methods were tested on glioblastoma image features for the precision of mIDH detection and compared to standard classification methods.

RESULTS: Using anomaly detection training methods, we were able to detect IDH mutations from features in noncontrast-enhancing regions in glioblastoma with an average precision of 75.0%, 69.9%, and 69.8% using three different models. Anomaly detection methods consistently outperformed traditional two-class classification methods from 2 unique learning models (67.9%, 67.6%). The disparity in performances could not be overcome through newer, popular models such as neural networks (67.4%).

CONCLUSION: We employed an anomaly detection strategy in the detection of IDH mutation in glioblastoma using preoperative T1 postcontrast imaging. We show these methods outperform traditional two-class classification in the setting of dataset imbalances inherent to IDH mutation prevalence in glioblastoma. We validate our results using an external dataset and highlight new possible avenues for radiogenomic rare event prediction in glioblastoma and beyond.

KEY WORDS: Glioblastoma, Radiomics, Machine learning

Neurosurgery 0:1–6, 2021 DOI:10.1093/neuros/nyab130 www.neurosurgery-online.com

Glioblastoma (GBM) is the most aggressive form of gliomas and the most common primary brain cancer in adults.[1] Molecular profiling has revealed distinct subtypes of GBMs that, though indistinguishable histologically, exhibit distinct natural history and therapeutic response.[2]. An important GBM subtype in this context involves those harboring mutations in the Isocitrate Dehydrogenase (IDH) gene.[3] While the physiologic function of IDH involves conversion of isocitrate to a-ketoglutarate,[4] cancer-causing mutations in IDH (mIDH) catalyze the conversion of isocitrate into D-2 hydroxyglutarate.[5] High levels of D-2 hydroxyglutarate induce a global alteration in the epigenome,[6] resulting in tumor physiology that is fundamentally distinct from those of wild-type IDH GBMs.[7] mIDH glioblastoma constitutes 5% to 13% of all GBMs.[8] Afflicted patients exhibit favorable prognosis,[9] particularly in patients who underwent gross total resections.[10] Thus,

ABBREVIATIONS: GAN, generative adversarial network; **GMM,** Gaussian mixture model; **IDH,** Isocitrate Dehydrogenase; **IHC,** immunohistochemistry; **mIDH,** Isocitrate Dehydrogenase mutated; **NGS,** next-generation sequencing; **OC,** one-class; **SVM,** support vector machine; **TCGA,** The Cancer Genome Atlas; **TCIA,** The Cancer Imaging Archive's; **UMMC,** University of Minnesota Medical Center

Supplemental digital content is available for this article at www.neurosurgery-online.com.

preoperative knowledge of the IDH mutation status bears an impact on surgical planning in the setting of newly diagnosed GBM.[11]

Magnetic resonance imaging (MRI) is the standard-of-care imaging modality for the diagnosis and management of GBM patients.[12] While the traditional role of MRI was limited to defining macroscopic tumor characteristics, including size, location, mass effect, and contrast enhancement,[12] emerging radiomic studies indicate that select MRI findings may serve as proxies for the molecular physiology of GBMs.[13] In this context, there is growing interest in identifying image features of MRI that define IDH mutations,[14] as well as developing machine learning models to predict IDH.[15,16] Efforts in the high accuracy diagnosis of IDH have been primarily driven by breakthroughs in deep artificial neural networks, often termed deep learning.[17] While the limited number of samples is a general challenge in medical image analytics, a particular challenge in applying machine learning to the detection of mIDH GBMs stems from the rarity of these tumors, effectively rendering them "outliers" in the statistical distribution of all GBMs. Treating the prevalent nonoutliers and the rare outliers as two classes and applying the typical classification methods on them lead to biased classification models that favor the dominant nonoutliers. Consequently, these models are unlikely to be useful in a real-world situation.

Here, we applied methods developed in the field of anomaly detection (also known as outlier detection) to detect mIDH GBM. In general, these methods strive to explicitly and effectively model the non-outliers as means to detect the outliers.[18] As an example, these methods are used by the financial industry to define normal patterns of customer spending in order to capture the anomalous, fraudulent transactions.[19] In this study, we defined mIDH GBM as anomalies within the distribution of all GBMs and utilized established anomaly detection methods for their detection. Expectedly, anomaly detection analytics performed better than the support vector machine (SVM) and neural network classification methods that we tested.

METHODS

Subjects and Genetic Data

Pre-operative, primary GBM scans were obtained from The Cancer Imaging Archive (TCIA)'s The Cancer Genome Atlas (TCGA) GBM collection. We subsetted this dataset to include samples with at least one T1 postcontrast scan, and IDH status obtained via next-generation sequencing (NGS) or immunohistochemistry (IHC). IDH mutational status was obtained from the TCGA Genomic Exploration Portal (portal.gdc.cancer.gov). In accordance with TCIA and TCGA policies, analysis of these samples do not require Institutional Review Board approval. An institutional Review Board-approved, retrospective analysis was conducted at the University of Minnesota Medical Center (UMMC) in order to identify World Health Organization (WHO) grade IV gliomas. As above, samples included for analysis required at least one T1-post contrast scan and IDH mutational status as determined via IHC or NGS.

Image Preprocessing; Feature Extraction

Brain extraction was done via automated methods.[20] Subsequently, N4 bias correction was used to correct for intensity nonuniformity.[21] We utilized the open-source radiomics toolbox, PyRadiomics, to extract image features based on morphology, intensity, and texture. Utilized features include: First Order Statistics, Shape-based 3D, Gray Level Co-Occurrence Matrix, Gray Level Run Length Matrix, Gray Level Size Zone Matrix, Neighbouring Gray Tone Difference Matrix, and Gray Level Dependence Matrix.

Region of Interest Segmentation

For both UMMC and TCGA cohorts, each patient's T1 postcontrast sequence was loaded into Slicer.[22] Due to limited availability of full MRI sequences in TCGA patients, T1 postcontrast images were selected in order to maximize sample size–as this was the most commonly available sequence. Tumor contrast-enhancing and noncontrast enhancing regions were segmented slice-by-slice in a semi-automatic fashion using Otsu thresholding.[23] All segmented regions were verified by visual inspection by a neurosurgery resident (B.T.).

Anomaly/Outlier Detection

Unlike traditional binary classification, which seeks to model both classes of data in order to classify future samples, anomaly detection methods seek to leverage a binary class imbalance through the modeling of the prevalent normal ("non-anomalous", "non-outlier") instances. In this manuscript, we define the prevalent IDH wild-type GBMs as our "normal" samples, and our mIDH as "anomalous".

One-Class SVM

The one-class (OC) SVM is an adaptation of the classic SVM for OC training in the setting of novelty or outlier detection. It works by constructing a maximally separating hyperplane that splits normal samples from outliers.[24]

Gaussian Mixture Model

A Gaussian mixture model (GMM) fits the nonoutliers using a parametric probability density function which consists of a mixture of Gaussian distributions. Statistical details of GMMs are found in **Supplemental Digital Content 1**.

Generative Adversarial Network

The generative adversarial network (GAN) involves a generator G and a discriminator D. The generator G learns to generate samples that resemble the true training samples, while the discriminator D, on the other hand, strives to tell apart the generated samples from the true samples.[25] After subsequent training, the discriminator is used to detect anomalies from normal samples. GANs have gained considerable attention in the field of deep learning in recent years.[26-28] In this paper, we leverage the distribution learning capability of GAN to explicitly model the distribution of the normal IDHwt GBM, which in turn helps us to single out the mIDH GBM as outliers. We save the details on GANs and our specific generator and discriminator architectures (they are realized as neural networks) in **Supplemental Digital Content 1**.

Training, Validation, and Testing Strategy

We implemented three train-test strategies: 1) Training and validating (model selection) on TCGA GBM samples, and testing on UMMC

TABLE. Baseline Patient Demographics of TCGA and UMMC

	TCGA	UMMC
Age	58.7	55.26
Male gender	60.1% (n = 113)	60.4% (n = 84)
IDH mutated % (n)	5.9% (n = 11)	12.3% (n = 17)
Race		
White	167	131
Black/African-American	11	3
Asian	0	2
Native American/American Indian	0	2
Not reported	7	0
Ethnicity		
Hispanic	2	0
Not Hispanic or Latino	148	138
Not reported	38	0

GBM samples, 2) Training and validating (model selection) on UMMC GBM samples, and testing on TCGA GBM samples, 3) Mixing both cohorts and training, validating, and testing on both cohorts.

Assessing Model Performance

In order to properly assess model performance in our data imbalanced setting, we used average precision (area under the precision recall curve) rather than accuracy. In imbalanced settings, accuracy (or the area under the receiver operating curve) has the potential to give misleading results.[29] We assess model robustness to training-validation-test splits by averaging model performance across 100 splits.

RESULTS

188 preoperative, T1 postcontrast images were obtained from patients with WHO Grade IV GBMs from the TCGA. 138 preoperative T1 postcontrast images were obtained from WHO Grade IV GBMs from UMMC. Table details patient demographic including age, gender, race/ethnicity, and IDH status. Baseline demographics between groups were significantly different in age (TCGA: 58.7; UMMC: 55.3; $P < .02$) but not in gender ($P = .173$). Age was added as a variable in our analyses. Potential baseline differences in race/ethnicity were not amenable due to lack of statistical power. Additionally, we mitigate potential confounding effects in two additional ways: iterative random splitting of training data and random mixing of training data from TCGA and UMMC.

Traditional Binary Classification Does Not Perform Well in mIDH-Scarce Settings in GBM

To establish a baseline performance for comparison, we used a traditional binary classification training scheme using an 80:20 training-test split (100 iterations). We trained a SVM and a random forest using radiomic features derived from non-contrast enhancing regions of tumors—in addition to patient age. Training on 188 TCGA samples, and tested on 138 UMMC samples. Using noncontrast enhancing regions, we were able to achieve an average precision of 67.9% and 67.6% using an SVM and random forest, respectively.

Deep Learning Models Cannot Overcome Class Imbalances

In order to explore more popular methods in radiogenomic binary classification, we trained a multi-layer neural network ("deep network") to explore whether more complex models can overcome learning imbalances. Training on TCGA GBM samples and testing on UMMC samples (100 iterations; 2-3 fully connected layers), we achieved an average precision of 67.4%.

Evaluating an Anomaly Detection Scheme for mIDH Imbalances in GBM

We trained a GAN, a GMM) and a OC-SVM using radiomic features derived from noncontrast enhancing regions of TCGA GBM samples, along with patient age, and tested on UMMC GBM samples. Using noncontrast enhancing regions, after 100 iterations, we were able to achieve an average precision of 75.0%, 69.9% and 69.8% using a GMM, GAN, and OC-SVM, respectively.

As Dataset Imbalance Worsens, Anomaly Detection Persistently Outperforms Traditional Binary Classification

In order to examine the effect of increasing testing imbalances, we varied the ratio of IDHwt/mIDH samples in our test data (while keeping the ratio in training data fixed). Using ratios extending from 1:2 to 1:10, GAN, GMM and OC-SVM consistently performed better than traditional binary classification (Figure 1).

Anomaly Detection Shows Robustness to Dataset-Dependent Effects

In order to explore any data-source dependent biases, we conducted 3 train-test strategies: 1) Training on TCGA and testing on UMMC; 2) Mixing samples from TCGA/UMMC for training and testing; 3) Training on UMMC and testing on TCGA. Using strategy (1), we report model performance using TCGA as training, and UMMC as testing data in the above sections. In strategy (2), samples were randomized to either the training or test set without respect to the data source (TCGA vs UMMC), we show these results on radiomics features with and without patient age included (Figures 2A and 2B). We achieve similar results on radiomics features with and without patient age in strategy (3) (Figures 3A and 3B). In addition, in all strategies (1-3), worsening imbalances in test data show anomaly detection consistently outperforms traditional two-class classification. These results are visualized in tabular format in **Supplemental Digital Content 2**.

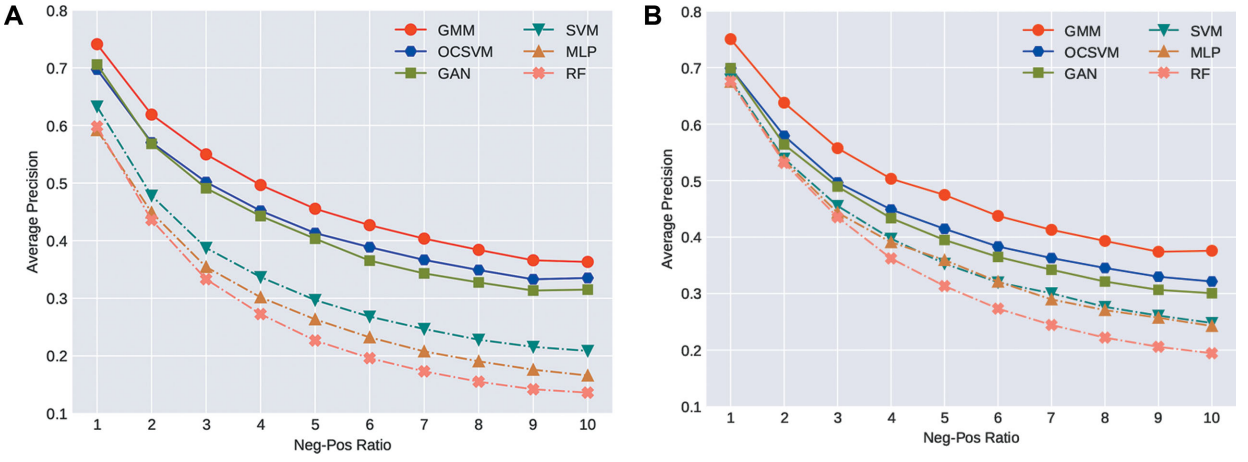

FIGURE 1. **A**, *Precision-recall curves for anomaly detection methods (GMM, OCSVM, GAN) as compared to traditional two-class classification methods (SVM, MLP, RF) using TCGA as training data and validating on unseen UMMC data.* **B**, *Precision-recall curves for abovementioned models when using age as a feature.*

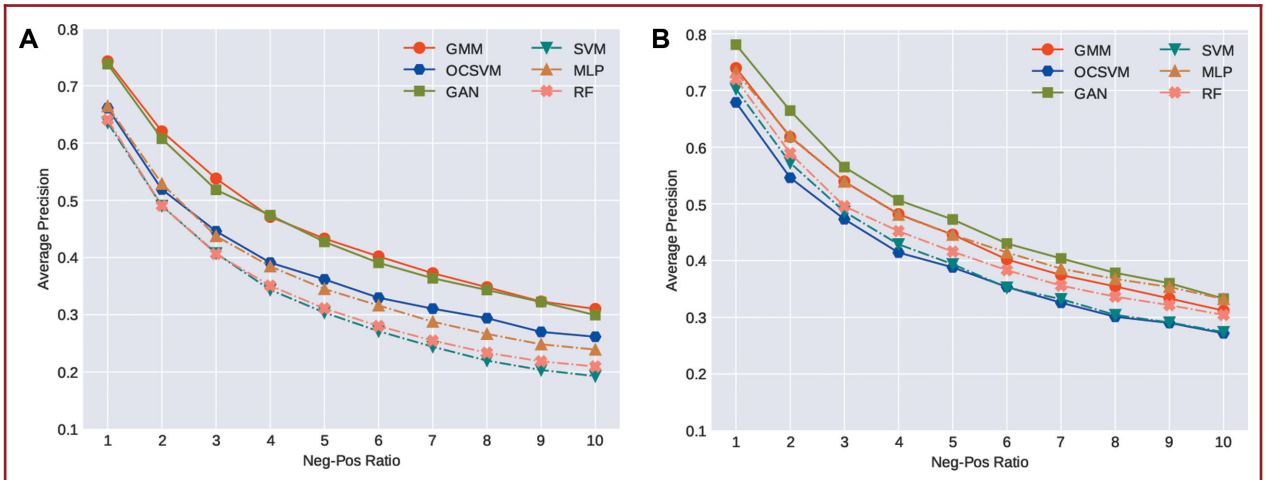

FIGURE 2. **A**, *Precision-recall curves for anomaly detection methods (GMM, OCSVM, GAN) as compared to traditional two-class classification methods (SVM, MLP, RF) using UMMC as training data and validating on unseen TCGA images.* **B**, *Precision-recall curves for abovementioned models when using age as a feature.*

DISCUSSION

Key Findings

In this work, we propose several models of anomaly/outlier detection in the identification of "rare-event" mutations of IDH in GBM. More specifically, we introduce 3 separate anomaly detection models: OC-SVM, GMM, and GANs as effective predictive tools in outlier detection. In our results, we show that virtually all anomaly detection models outperform traditional two-class classification models. In particular, GMMs with fewer components (K = 1-4) have the greatest performance in mIDH detection in GBM—outperforming more complicated neural network style models. In addition, we show that as the data imbalance worsens, the benefits of anomaly detection training become more accentuated. We also reviewed the potential of more cutting edge tools like GANs, which have recently found utility in numerous applications including synthetic data generation.[25,26] We present evidence that while anomaly detection GANs do certainly outperform traditional two-class classification models, they do not outperform other, simpler anomaly detection parametric models like GMMs, which may present evidence of diminished utility of GANs in the small data context and/or

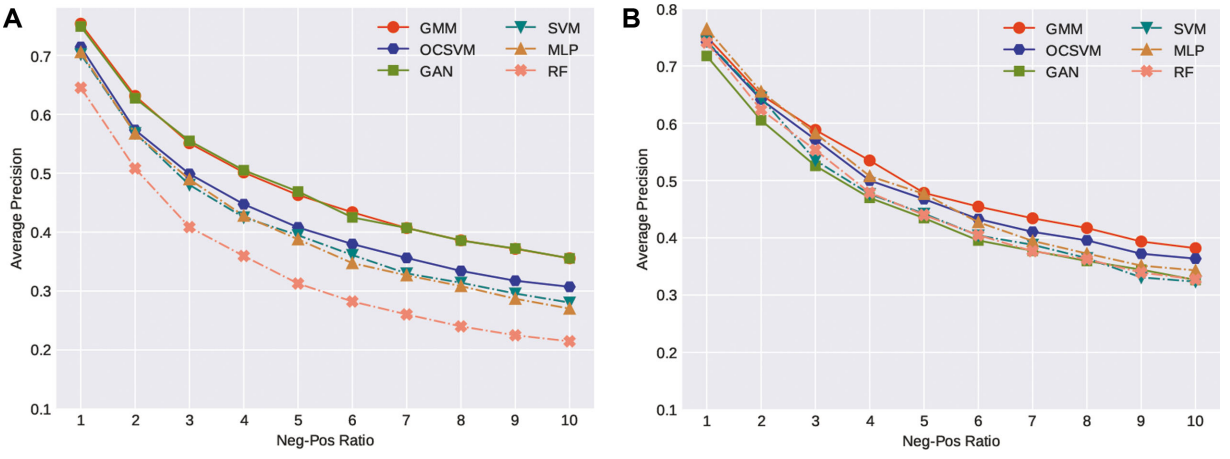

FIGURE 3. **A**, *Precision-recall curves for anomaly detection methods (GMM, OCSVM, GAN) as compared to traditional two-class classification methods (SVM, MLP, RF) using a mix of TCGA and UMMC as training data and validating on unseen TCGA and UMMC images.* **B**, *Precision-recall curves for abovementioned models when using age as a feature.*

anomaly detection setting. However, more data is required to validate this hypothesis. While GMMs do show promise in this setting, with the acquisition of more data, the capabilities of GANs may be better realized.

Interpretation

Our study highlights the danger of blind reliance on machine learning and deep learning techniques without careful study of the underlying model and balanced representation of inputted data. Improper application of conventional techniques in imbalanced data settings may lead to false conclusions given the misleading, inflated overall performance. Recently, a new movement in "equitable machine learning" has sparked interest in investigating the potential perils and consequences of commercial-grade predictive models for hiring/firing, prison sentences, and even loan approvals that have been trained on imbalanced data.[30,31]

Limitations

Several limitations exist in our study. Our study assumes inputted samples have a diagnosis of GBM–something a clinician would not know for certain before surgery. The ideal, deployed model would exist in an end-to-end fashion (from histological type to grading to genotyping). However, the paucity of data severely limits the ability to construct this with one single model but rather as a sequence of predictive models in series. Future studies to tackle the end-to-end problem would leverage the power of individual models involving grading only or genotyping. Our study sample size (n = 326), while more than most radiomic studies in glioma, is less than many other non-oncological medical imaging studies. We mitigate this relatively small sample size by limiting our feature space to 107 features. Furthermore, in order to adequately reduce sample complexity, we do not include imaging from other sequences (T1 precontrast, T2, FLAIR). Future studies can leverage multiple sequences while at the same time balancing problems in overfitting through feature selection techniques.

Generalizability

In the context of general applicability (particularly clinical applicability), predictive models are expected to perform consistently and equitably before deployment in practice. These expectations are solely possible given unbiased, adequate representation in training data, or specialized tools/training strategies which address these issues. In the radiogenomics setting, particularly in GBM, the accumulation of representative training data in mIDH is an epidemiologically challenging task that cannot be addressed by classical tools in machine learning.

CONCLUSION

An anomaly detection paradigm represents a novel approach in leveraging known information to better detect rare events. This tool represents an early, but major step in preoperative identification of mIDH with the potential for significant implications in surgical planning.

Funding

This study did not receive any funding or financial support.

Disclosures

The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article.

REFERENCES

1. Dolecek TA, Propp JM, Stroup NE, Kruchko C. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005-2009. *Neuro Oncol*. 2012;14(Suppl 5):v1-v49.
2. Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110.
3. Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 mutations in gliomas. *N Engl J Med*. 2009;360(8):765-773.
4. Yang H, Ye D, Guan K-L, Xiong Y. IDH1 and IDH2 mutations in tumorigenesis: mechanistic insights and clinical perspectives. *Clin Cancer Res*. 2012;18(20):5562-5571.
5. Flavahan WA, Drier Y, Liau BB, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*. 2016;529(7584):110-114.
6. Sturm D, Witt H, Hovestadt V, et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell*. 2012;22(4):425-437.
7. Raineri S, Mellor J. IDH1: linking metabolism and epigenetics. *Front Genet*. 2018;9:493.
8. Williams Parsons D, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*. 2008;321(5897):1807-1812.
9. SongTao Q, Lei Y, Si G, et al. IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci*. 2012;103(2):269-273.
10. Beiko J, Suki D, Hess KR, et al. IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. *Neuro Oncol*. 2014;16(1):81-91.
11. Almeida JP, Chaichana KL, Rincon-Torroella J, Quinones-Hinojosa A. The value of extent of resection of glioblastomas: clinical evidence and current approach. *Curr Neurol Neurosci Rep*. 2015;15(2):517.
12. Pope WB, Brandal G. Conventional and advanced magnetic resonance imaging in patients with high-grade glioma. *Q J Nucl Med Mol Imaging*. 2018;62(3):239-253.
13. Fathi Kazerooni A, Bakas S, Saligheh Rad H, Davatzikos C. Imaging signatures of glioblastoma molecular characteristics: a radiogenomics review. *J Magn Reson Imaging*. 2020;52(1):54-69.
14. Li Z-C, Bai H, Sun Q, et al. Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma. *Cancer Med*. 2018;7(12):5999-6009.
15. Liu X, Li Y, Li S, et al. IDH mutation-specific radiomic signature in lower-grade gliomas. *Aging*. 2019;11(2):673-696.
16. Lu C-F, Hsu F-T, Hsieh KL-C, et al. Machine learning-based radiomics for molecular subtyping of gliomas. *Clin Cancer Res*. 2018;24(18):4429-4436.
17. Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin Cancer Res*. 2018;24(5):1073-1081.
18. Abdel-Aziz AS, Hassanien AE, Azar AT, Hanafi SE-O. Machine learning techniques for anomalies detection and classification. In: *Advances in Security of Information and Communication Networks*. Springer Berlin Heidelberg; 2013:219-229.
19. Niu X, Wang L, Yang X. A comparison study of credit card fraud detection: supervised versus unsupervised. 2019. http://arxiv.org/abs/1904.10604. Accessed May 1, 2020.
20. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp*. 2019;40(17):4952-4964.
21. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320.
22. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-1341.
23. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst, Man, Cybern*. 1979;9(1):62-66.
24. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Comput*. 2001;13(7):1443-1471.
25. Goodfellow I, Pouget-Abadie J, Mirza M. Generative adversarial nets. *Adv Neural Inf Process Syst*. 2014. http://papers.nips.cc/paper/5423-generative-adversarial-nets. Accessed June 2, 2020.
26. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. 2016. http://arxiv.org/abs/1607.02533. Accessed February 20, 2020.
27. Truong ND, Zhou L, Kavehei O. Semi-supervised seizure prediction with generative adversarial networks. *Conf Proc IEEE Eng Med Biol Soc*. 2019;7:2369-2372.
28. Zhao M, Liu X, Liu H, Wong KKL. Super-resolution of cardiac magnetic resonance images using Laplacian Pyramid based on Generative Adversarial Networks. *Comput Med Imaging Graph*. 2020;80(Mar):101698.
29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
30. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London, UK: Crown; 2016.
31. Eubanks V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York City, NY, USA: St. Martin's Publishing Group; 2018.

Supplemental digital content *is available for this article at www.neurosurgery-online.com.*

Supplemental Digital Content 1. Methods.

Supplemental Digital Content 2. Table. A, Tabular representation of precision-recall performance using TCGA as training data and testing on UMMC data. B, Tabular representation of precision-recall performance using UMMC as training data and testing on TCGA data. C, Tabular representation of precision-recall performance using a mix of UMMC and TCGA as training data and testing on unseen TCGA and UMMC data.